SoGrIn: a non-verbal dataset of social group-level interactions

Nicola Webb[1], Manuel Giuliani[1] and Séverin Lemaignan[2]

Abstract— **We present the Social Group Interactions (SoGrIn) dataset; a dataset which captures non-verbal signals of groups as they complete socially collaborative and formation-provoking tasks. The dataset comprises precise proxemics (captured motion) and facial features (facial landmarks, gaze direction, facial action units) encompassing a total duration of 60 minutes involving 30 individuals, divided into six groups. Also included are basic demographic information and responses to the Big 5 personality questionnaire. The Social Group Interactions dataset is publicly available at `https://doi.org/10.5281/zenodo.7778123`.**

I. Introduction

When humans come together to start an interaction, there is an exchange of both verbal and non-verbal social signals to establish our presence and intentions. We continue to share these signals throughout the interaction to portray our thoughts and motivations. As humans, we can interpret these signals easily, requiring little conscious cognitive effort. Social robots, however, are not unable to decipher this information so easily.

To do this, social robots must be able to take in and learn from vast amounts of human social behaviour data. By using a data-driven approach, robots will be able to recognize patterns and interpret social signals in a way that aligns with our own. This is important for enabling robots to engage with us with more natural and intuitive behaviour, and also to ensure that they can navigate complex social situations with ease and sensitivity.

In this paper, we present the Social Group Interactions (SoGrIn) dataset. We describe the methodology for provoking group interactions and the technologies used to collect this data. This dataset fills a gap in the existing publicly available datasets of human group behaviours, which have been limited in their ability to capture the complexity and dynamics of real-life social interactions.

A. Dataset applications

We see the SoGrIn dataset having value in the domain of social signal understanding. One of the main challenges in this area concerns the multi-modal nature of our communication. For social robots, by combining the extraction and analysis of multiple complementary signals, we can create a framework that can build over time a picture of what is occurring in an interaction. For example, if a robot can hear several humans talking, but cannot visibly see the humans, it can deduce that there are at least as many humans in that location as humans are talking. With the addition of facial detection, the robot can now place each human that it visibly detects in its environment. By using multimodal learning, we can execute continuous detection even with missing data due to the ever-changing dynamics of social interaction.

To achieve this level of sophistication in social robots, large amounts of real-world data are needed. In this context, the dataset we have collected can be a valuable resource for creating *data-driven* models to detect and interpret social cues in the context of human-robot interactions.

In addition, the diverse set of signals that have been collected in this dataset makes it possible to model and analyze both the complex *social dynamics* that occur during human-human and human-robot interactions. The data capture both human-human and human-robot interactions, providing a unique opportunity to study and understand the intricacies of social behaviour in these situations.

Finally, our motion capture data allows for the precise measurement of proximity during human-robot interactions. This information can be utilized for *proximity-based analysis* of social communication.

II. Related Work

Several datasets are publicly available in the domain of social signal processing, including Facial Emotion Recognition (FER) [10], speech understanding [13] and person detection [4]. As such, the collection of social multimodal datasets is not a new concept. Many of these traditional datasets are however focused on simple tasks with limited (or excessively constrained) social interactions.

Looking at natural human social interaction, we have performed an extensive review of existing multi-modal datasets, and found a range of datasets that captures a diverse set of group interactions, with a range of tasks and group sizes.

The UoL-3D Social Interaction dataset [6] consists of 20 videos containing 10 pairs of participants. Participants were asked to act out various individual activities, such as making a cup of coffee, whilst periodically being engaged in social interaction with another participant.

The Synergetic sociAL Scene Analysis (SALSA) dataset [1] consists of 60 minutes of footage from two social events, a poster presentation and a cocktail party. The data includes video and audio recordings and is annotated with head and body orientation along with the F-formations.

The Idiap Wolf dataset [8] contains video and audio recordings from 15 rounds (36 participants) of the role-playing game 'Werewolf', for a total of about 81 hours.

[1] Nicola Webb and Manuel Giuliani are with Bristol Robotics Lab, University of the West of England, Bristol, United Kingdom, `firstname.lastname@brl.ac.uk`

[2] Séverin Lemaignan is with PAL Robotics, Barcelona, Spain `severin.lemaignan@pal-robotics.com`

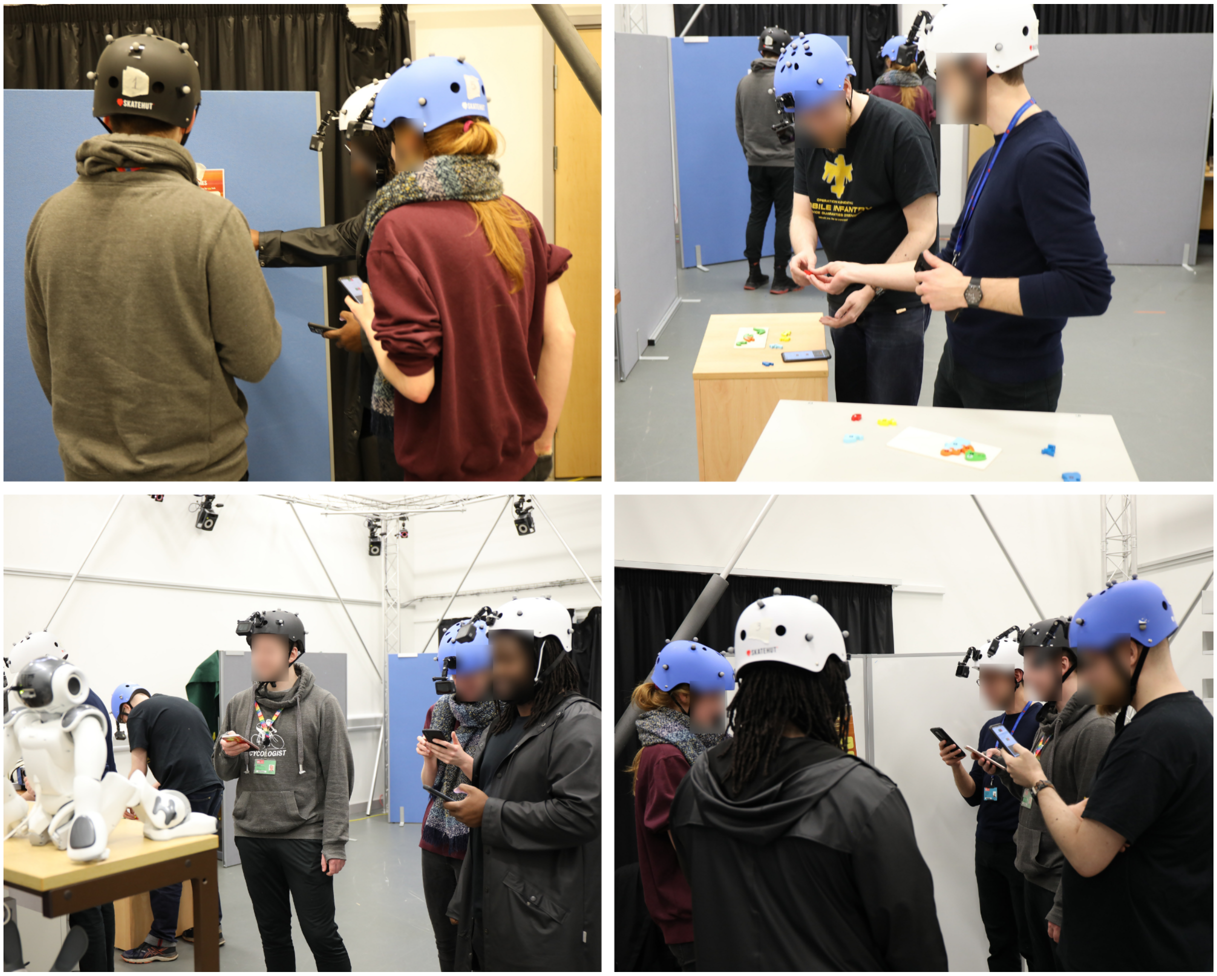


Fig. 1. Image showing participants completing tasks in each of the 4 'rooms'. (top left) Discussion task in room 1 (top right) Collaborative jigsaw completion task in room 2 (bottom left) Human-robot interaction task in room 3 (bottom right) Debate task in room 4.

The Multimodal Focused Interaction dataset [2] is a first-person perspective approach to an interaction dataset. It contains 19 videos (no audio) showing focused and unfocused social interaction in the first person.

The AMI corpus [5] consists of 100 hours of video and audio recordings from meetings that are either staged or genuine. Annotations of head and hand gestures are provided, as well as gaze direction, participants' movements around the meeting room, emotional state and the locations of heads in video frames. The corpora was one of the earliest fully annotated and transcribed datasets in the field.

The D64 corpus [12] consists of 8 hours of audio and video footage from 2 unscripted conversations between 4 to 5 participants. 7 cameras were employed, with a camera focused on each participant. Two further 360° cameras were used to record the entire interaction. The participant pool is university professors with several master's students. This provides the opportunity for an investigation into social dynamics between people within a hierarchy.

The PInSoRo dataset [9] consists of 45 hours of social play between 45 child-child pairs and 30 child-robot pairs. The dataset is highly annotated and includes facial landmarks, action units, head pose estimation, gaze estimation, 2D skeleton data (body and hand tracking), audio, and annotations of timestamped social behaviours and events.

The CongreG8 dataset [16] consists of 418 recordings of free-standing groups, with and without a robot, and specifically look into group joining: an external participant (or robot) joins the group, and the dataset records the resulting group motions using motion capture.

The GAME-ON dataset [11] consists of 11 hours of footage of small groups playing a social 'escape' themed game in a lab environment. The groups were made up of 3 people who had a pre-existing friendship. Groups were asked to complete 5 tasks that each promoted a varying level of collaboration from group members to record different levels of cohesion. Participants were fitted with motion tracking suits for full-body recording.

The Unobtrusive Group Interaction (UGI) Corpus [3] is a multimodal dataset consisting of audio and video recordings from group meetings. Participants were given a collaborative group task to complete for roughly 15 minutes and in total 22 meetings were recorded.

The TED Gesture Dataset [17] consists of 1295 videos

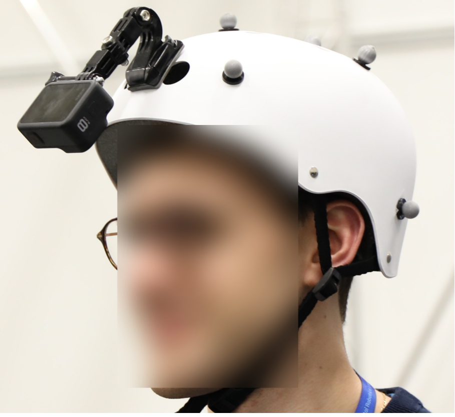


Fig. 2. Helmet worn by participants showing GoPro attachment and vicon markers.

taken of numerous TED talks where one or more speakers give speeches for roughly 10 minutes. Videos were segmented to capture only the frames where the speaker was forward facing and where either their upper or entire body was visible. The dataset does not directly include interaction data.

Finally, the MUMBAI dataset [7] consists of video and audio recordings of groups of 4 participants playing various cooperative board games. 62 sessions were recorded, with each game varying in length.

While this review shows that numerous datasets are already available, we observe that it is difficult to find datasets that provide accurate position and orientation of participants (using eg motion capture) for natural social interaction, like the ones likely to be encountered by a social robot interacting in a human environment.

The GAME-ON dataset [11] is probably the most similar to our contribution. Compared to GAME-ON, and while smaller in size and focusing on non-audio cues only, our dataset includes a broader range of social situations, as well as detailed facial information while the participants interacted. Furthermore, due to the nature of the individual tasks, we intentionally stimulate interactions beyond the scope of regular group tasks, generating a broader range of interactions. As such, the range and diversity of naturalistic group interaction found in our dataset is a superset of the situations recorded in GAME-ON, and provides support for new research venues on social dynamics.

III. Methodology

A. Game Protocol

To promote dynamic group interactions, each session was designed as a short competitive game. The game involved completing a series of tasks within a 10-minute timeframe. These tasks were divided into two sets - one to be completed by the entire group, and the other to be attempted individually while keeping it a secret from other participants. Some examples of individual tasks: make someone laugh, form a group with 2 people, get two people to talk to each other, bring one person to room X, speak to someone you haven't spoken to yet. The group tasks could be accomplished by the whole group or smaller breakout ones.

Points were awarded for successfully completing a task, with both group and individual tasks being worth 3 points each. However, failing or skipping a task resulted in a deduction of 3 points. The participants were encouraged to aim for the highest score to 'win' the game. To maintain the competitive spirit, those who were caught attempting individual tasks were marked as 'failed.'

All tasks had predominantly social undertones, such as 'complete a jigsaw together'. Group tasks predominantly directed participants to a room, in which they would have to pick a specific task to do from a list. All possible tasks are shown in Table I. The individual tasks were focused on trying to manipulate the other participants, either by getting them to do a different task than planned or by distracting someone from what they were doing. Furthermore, we present the various types of interactions observed during the completion of these tasks. In cases where multiple interaction types were identified, it signifies that these interaction types were witnessed at least once.

Participants each had a helmet, shown in Figure 2. Each helmet was fitted with a GoPro camera on a mount positioned at the participant's face. They were to keep this on for the duration of the data recording.

At the beginning of the study, participants were directed to a Web App on their mobile phones that they were to use for the duration of the study. On the Web App, both their basic demographic information; *helmet number, age, the culture you identify with, gender* and responses to the Short IPIP-BFM-20 Questionnaire [14] were collected.

All tasks were delivered through the same web application, as described in [14]. To mark the completion of a task, participants had to click on a green tick icon, while a red cross icon was used to indicate task failure or skipping. This web application was designed to simplify the task completion process and facilitate data collection.

TABLE I

Possible tasks in each room.

Room	Tasks	Interactions
1	Quick verbal game: 20 Questions, Rhyming Chain, True or False	1-1, 1-N, N-N
2	Complete both jigsaws	1-1, 1-N, N-N
3	Tell NAO a joke Give NAO a new name Guess NAO's favourite number	1-N
4	Debate a topic for 1 minute	N-N

The experimental space was divided into four distinct rooms, each designed to elicit different forms of group interaction. Participants were directed to each room in turn to complete specific tasks: Room 1 focused on discussions, Room 2 on collaborative tasks, Room 3 on robot interactions, and Room 4 on debates. This setup encouraged participants to move around the space and engage in diverse forms of social interaction. By varying the types of tasks and the nature of the interaction required in each room, the experiment sought to capture a wide range of social signals and behaviours.

The key social situations that are frequently captured during recording are as follows: people making each other laugh, smiling at each other, working together on a task, introducing oneself to a group, talking to a NAO robot, using a mobile phone and leading a group to a location.

B. Social Signal Capture

1) VICON: The Vicon Vero motion capture system was used for recording participant location in the space and their head orientation. 6 cameras were placed around the room to capture as much as possible and reduce occlusions. Bike helmets were used to attach markers on each participant, shown in Figure 3. Nine VICON makers were attached to each helmet, spaced apart to reduce the number of camera occlusions.

The plot depicted in Figure 4 represents the motion capture recording of participants during a single session, with each room being highlighted for clarity.

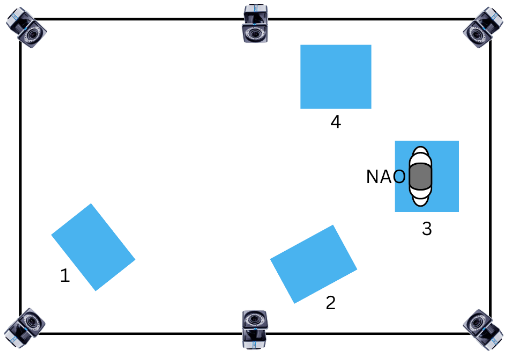


Fig. 3. Top-down view of the experimental space, showing locations of the vicon motion capture cameras and an estimated position of where the tasks rooms were located.

2) GoPro: Each helmet was fitted with a GoPro HERO 8, positioned to record the participant's face.

In Figure 5, we show an example of how AUs can be detected in a snapshot from one of the GoPro recordings. While FACS and AU detection have been used extensively in research on facial expression and emotion recognition, there is still much work to be done to improve the accuracy and reliability of these methods. As facial expressions are just one aspect of social signal processing, we must explore other signals such as speech, gesture, and body language to better understand the complex dynamics of social interactions.

3) NAO: Python: For tasks including interacting with the NAO, a Wizard of Oz approach was used. When participants spoke to the robot, a pre-written statement was manually selected in response, using a python script.

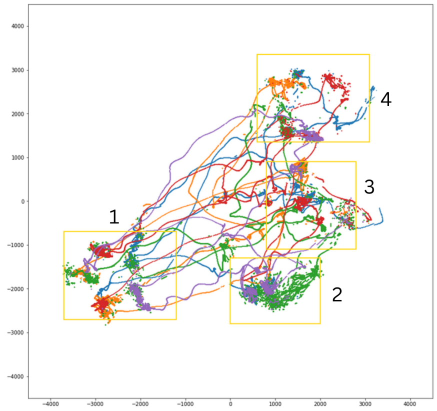


Fig. 4. Plot showing motion capture recording of participants from one session. Each room has been highlighted.

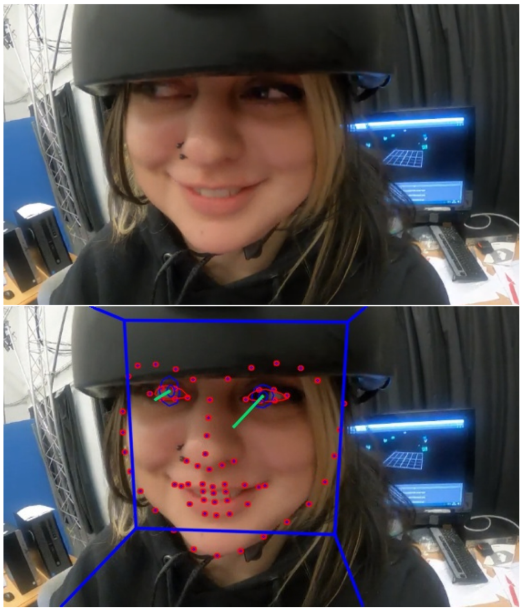


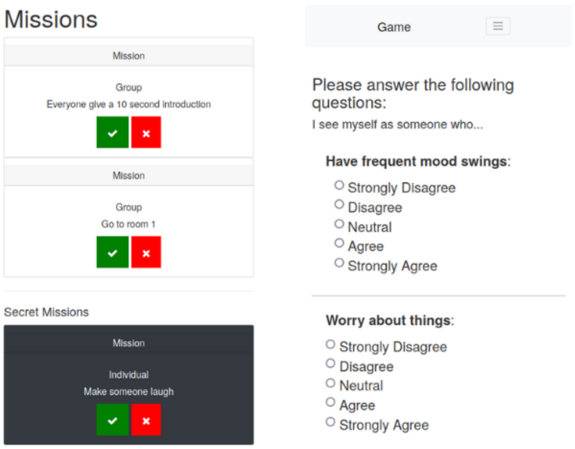
Fig. 5. Image showing action unit detection of one frame of GoPro video.

4) Web App: Flask: To avoid encountering platform-specific requirements, Flask was used to develop the web app. Figure 6 shows two pages of the application, including the missions page where tasks are assigned and the Big 5 Questionnaire.

C. Participants

30 participants were recruited campus-wide at the University of the West of England. Participants were aged 24-59 (SD=8.8, M=30) with a gender split of 20 males, 9 females and 1 other. They were recruited into 6 groups of 5 based on their availability. They were asked to bring their own mobile phones to the study to access the Web App. They were compensated with a £5 Amazon voucher for their time.

1) Big 5 Questionnaire: All 30 participants' responses to the Big 5 questionnaire are represented in Figure 7, which displays the distribution of these responses. This



Missions

Mission

Group

Everyone give a 10 second introduction

✓ ✗

Mission

Group

Go to room 1

✓ ✗

Secret Missions

Mission

Individual

Make someone laugh

✓ ✗

Game

Please answer the following questions:

I see myself as someone who...

Have frequent mood swings:

○ Strongly Disagree
○ Disagree
○ Neutral
○ Agree
○ Strongly Agree

Worry about things:

○ Strongly Disagree
○ Disagree
○ Neutral
○ Agree
○ Strongly Agree

Fig. 6. Snapshot of two pages of the Web App (left) Showing the task/'missions' page (right) Showing the Big 5 questionnaire

information can be used to gain insights into the participants' personalities, as the Big 5 questionnaire is a widely used tool for assessing individuals' traits across the five dimensions of openness, conscientiousness, extraversion, agreeableness, and neuroticism.

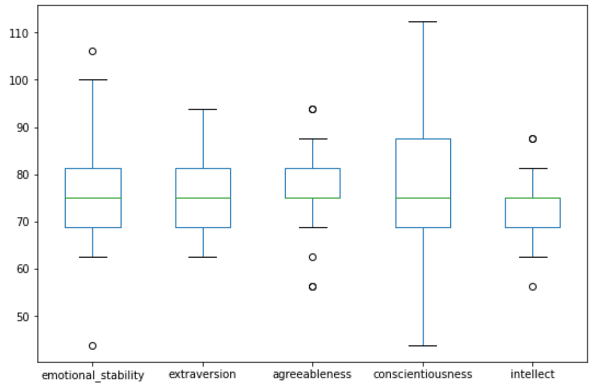


Fig. 7. Graph showing the Big 5 personality questionnaire responses and the deviation for each

IV. Conclusion

In order to address a gap in the existing, publicly-available, datasets of human group behaviours, we have designed a protocol and recorded a novel dataset of realistic human social behaviours, with a focus on group activities.

The dataset includes recordings of 30 people (spread over 6 groups of 5 persons) while participating to a playful group activity. The game comprises of several mini-tasks (*missions*) that need to be performed by the players in group (like 'debate topic X together'), as well as a small set of 'covert' individual missions (like 'make player X laugh'), designed to create a variety of social situations, while remaining engaging and fun to elicit as natural as possible behaviours.

We recorded accurate proxemics via motion capture, as well as facial features (like facial action units and gaze direction) for each participants, for a total duration of approximately 60 minutes.

The resulting dataset, the *Social Group Interactions* dataset, is publicly available at `https://doi.org/10.5281/zenodo.7778123`.

A. Main limitations

Firstly, the data was collected in a lab setting. Despite designing our game protocol to inspire more natural interactions, the participants were still aware that they were being recorded, which may have influenced their behaviour in some way. Despite efforts to create a more natural and engaging environment for participants through the game protocol, the awareness of being recorded may have still affected the authenticity of their interactions. Therefore, the SoGrIn dataset should be used with an understanding of its limitations in capturing fully natural social behaviours.

Another limitation of the SoGrIn dataset is that the facial recordings were not captured from the perspective of the participants. This means that researchers do not have access to information about what participants could see during the interactions, which may have affected their behaviour. Although the dataset does capture facial expressions, this limitation should be considered when interpreting the data.

Furthermore, it is worth noting that the SoGrIn dataset lacks audio recordings, which could have offered valuable insights into the nuances of verbal communication and the tone of voice employed during social interactions. Nevertheless, the GoPro footage does enable us to determine whether a participant is engaged in conversation or not. Although the GoPros did record audio, it is important to acknowledge that the audio quality falls short, thereby posing challenges for accurate transcription.

Finally, the overall size of the dataset (1 hour) is relatively small compared to other publicly available datasets. This limitation should be considered when planning research studies that require large amounts of data.

Despite these limitations, the SoGrIn dataset still offers a valuable resource for those interested in studying social group interactions. By acknowledging these limitations, the dataset can be used in a thoughtful and informed way to gain insights into social behaviour within groups.

B. Future work: comparison with social engagement metrics in virtual environment

In previous work [15], we completed a data collection during the covid pandemic, using an online game where participants played an online socially interactive game, simulating simple group interactions. From this data, we created the visual social engagement metric, a metric designed to detect someone's engagement level using only two visual social signals: interpersonal proximity and mutual gaze. We look to enrich our metric by using the data collected in this work by first using real-world data instead of simulation and secondly by adding additional signals to the metric, such as the facial point data mentioned previously. With the knowledge of how engaged someone in an interaction is, it is possible to make an educated decision on the next steps for one's ways of interacting.

Acknowledgement

This work has received funding from the European Union's H2020 ACCIO TecnioSpring INDUSTRY project (grant agreement no. 801342, project TALBOT). Funding also received from the Wallscourt Foundation.

References

[1] Alameda-Pineda, X., Staiano, J., Subramanian, R., Batrinca, L. M., Ricci, E., Lepri, B., Lanz, O., and Sebe, N. Salsa: A novel dataset for multimodal group behavior analysis. *CoRR abs/1506.06882* (2015).

[2] Bano, S., Suveges, T., Zhang, J., and McKenna, S. Multimodal egocentric analysis of focused interactions. *IEEE Access 6* (June 2018), 37493–37505.

[3] Bhattacharya, I., Foley, M., Ku, C., Zhang, N., Zhang, T., Mine, C., Li, M., Ji, H., Riedl, C., Welles, B. F., et al. The unobtrusive group interaction (ugi) corpus. In *Proceedings of the 10th ACM Multimedia Systems Conference* (2019), pp. 249–254.

[4] Borges, P. V. K., Conci, N., and Cavallaro, A. Video-based human behavior understanding: A survey. *IEEE Transactions on Circuits and Systems for Video Technology 23*, 11 (2013), 1993–2008.

[5] Carletta, J. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation 41*, 2 (2007), 181–190.

[6] Coppola, C., Cosar, S., Faria, D., and Bellotto, N. Automatic detection of human interactions from rgb-d data for social activity classification. In *IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN)* (2017), pp. 871–876.

[7] Doyran, M., Schimmel, A., Baki, P., Ergin, K., Türkmen, B., Salah, A. A., Bakkes, S. C., Kaya, H., Poppe, R., and Salah, A. A. Mumbai: multi-person, multimodal board game affect and interaction analysis dataset. *Journal on Multimodal User Interfaces 15*, 4 (2021), 373–391.

[8] Hung, H., and Chittaranjan, G. The idiap wolf corpus: Exploring group behaviour in a competitive role-playing game. In *Proceedings of the 18th ACM International Conference on Multimedia* (New York, NY, USA, 2010), Association for Computing Machinery, p. 879–882.

[9] Lemaignan, S., Edmunds, C. E., Senft, E., and Belpaeme, T. The pinsoro dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PloS one 13*, 10 (2018).

[10] Li, S., and Deng, W. Deep facial expression recognition: A survey. *CoRR abs/1804.08348* (2018).

[11] Maman, L., Ceccaldi, E., Lehmann-Willenbrock, N., Likforman-Sulem, L., Chetouani, M., Volpe, G., and Varni, G. Game-on: A multimodal dataset for cohesion and group analysis. *IEEE Access 8* (2020), 124185–124203.

[12] Oertel, C., Cummins, F., Edlund, J., Wagner, P., and Campbell, N. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces 7*, 1-2 (2013), 19–28.

[13] Serban, I. V., Lowe, R., Henderson, P., Charlin, L., and Pineau, J. A survey of available corpora for building data-driven dialogue systems, 2017.

[14] Topolewska-Siedzik, E., Skimina, E., Strus, W., Cieciuch, J., and Rowiński, T. The short ipip-bfm-20 questionnaire for measuring the big five. *Roczniki Psychologiczne // Annals of Psychology 17* (01 2014), 385–402.

[15] Webb, N., Giuliani, M., and Lemaignan, S. Measuring visual social engagement from proxemics and gaze. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (2022), IEEE, pp. 757–762.

[16] Yang, F., Gao, Y., Ma, R., Zojaji, S., Castellano, G., and Peters, C. A dataset of human and robot approach behaviors into small free-standing conversational groups. *PLOS ONE 16* (02 2021), 1–24.

[17] Yoon, Y., Ko, W.-R., Jang, M., Lee, J., Kim, J., and Lee, G. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)* (2019), pp. 4303–4309.