A Multimodal Dataset for Robot Learning to Imitate Social Human-Human Interaction

Nguyen Tan Viet Tuyen
Department of Engineering
King's College London
London, United Kingdom
tan_viet_tuyen.nguyen@kcl.ac.uk

Alexandra L. Georgescu
Department of Psychology
King's College London
London, United Kingdom
alexandra.livia.georgescu@gmail.com

Irene Di Giulio
Centre for Human & Applied Physiological Sciences
King's College London
London, United Kingdom
irene.di_giulio@kcl.ac.uk

Oya Celiktutan
Department of Engineering
King's College London
London, United Kingdom
oya.celiktutan@kcl.ac.uk

ABSTRACT

Humans tend to use various nonverbal signals to communicate their messages to their interaction partners. Previous studies utilised this channel as an essential clue to develop automatic approaches for understanding, modelling and synthesizing individual behaviours in human-human interaction and human-robot interaction settings. On the other hand, in small-group interactions, an essential aspect of communication is the dynamic exchange of social signals among interlocutors. This paper introduces LISI-HHI – Learning to Imitate Social Human-Human Interaction, a dataset of dyadic human interactions recorded in a wide range of communication scenarios. The dataset contains multiple modalities simultaneously captured by high-accuracy sensors, including motion capture, RGB-D cameras, eye trackers, and microphones. LISI-HHI is designed to be a benchmark for HRI and multimodal learning research for modelling intra- and interpersonal nonverbal signals in social interaction contexts and investigating how to transfer such models to social robots.

CCS CONCEPTS

• **Human-centered computing → User models**.

KEYWORDS

human-human interaction, multimodal dataset, nonverbal behaviour analysis and synthesis

ACM Reference Format:

Nguyen Tan Viet Tuyen, Alexandra L. Georgescu, Irene Di Giulio, and Oya Celiktutan. 2023. A Multimodal Dataset for Robot Learning to Imitate Social Human-Human Interaction. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3568294.3580080

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03…$15.00
https://doi.org/10.1145/3568294.3580080

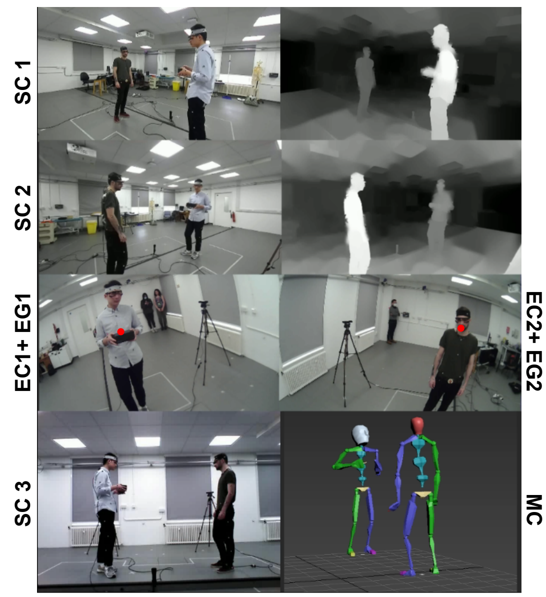


Figure 1: Multimodal data featuring in the LISI-HHI dataset. (SC1: Statistic RGB-D 1, SC2: Statistic RGB-D 2, EC1: Egocentric RGB 1, EG1: Eye gaze of person P1, EC2: Egocentric RGB 2,EG2: Eye gaze of person P2, SC3: Statistic RGB D 3, MC: motion data)

1 INTRODUCTION

Nonverbal behaviours play essential roles in social interaction. Humans tend to use various social signals, including facial expressions, body gestures, eye gaze, and vocal expressions, to communicate their messages to interaction partners. Vice versa, they interpret social cues observed from others to understand the interaction context better. Previous studies utilised such social signals as essential clues to develop automatic approaches for modelling user engagement [5], emotion [24], intention [20], personality [22] as well as synthesizing individual behaviours [9, 29] in both human-human interaction (HHI) and human-robot interaction (HRI) settings.

On the other hand, in small-group interaction, an essential aspect of communication is the dynamic exchange of nonverbal signals among interlocutors, with the aim of adapting to social norms [15] and building a common ground [21]. This social factor suggests that both intra- and interpersonal social signals should be considered when modelling the interlocutors' behaviours. The idea has been implemented in recent works for modelling individual personality [23] as well as synthesising facial expressions and body gestures in human-agent or human-robot interaction [12, 27, 28]. Capturing both intra- and interpersonal social signals enables the learning framework to better model the interaction context [12, 23, 27]. Consequently, this paper introduces LISI-HHI, a multimodal dataset of dyadic interactions, to foster the development of recent efforts in this potential research direction. LISI-HHI consists of multiple nonverbal modalities captured simultaneously via multiple high-accuracy sensors (e.g. motion capture system, and eye-tracker system). Fig. 1 presents an example from the proposed dataset. To the best of our knowledge, LISI-HHI is among a few available databases that cover a high number of modalities, camera views, participants, and interaction sessions.

2 AVAILABLE HHI DATASETS

Table 1 summarizes publicly available datasets in this domain. The comparison of available databases, in terms of modalities, are illustrated in Table 2. We only review dyadic datasets that contain at least an audio and a visual channel.

The LISI-HHI dataset complements the previous databases by incorporating a multi-sensory setup with a novel design of multiple interaction scenarios. Without estimating skeleton data from RGB images – as implemented in earlier works [1, 4, 23], we used a motion capture system to capture interlocutors' motions at a high frequency. The experiment was set up with eye-trackers to ensure the accuracy of eye gaze data collected, similarly to previous works [20, 25]. In contrast with the other motion capture datasets [2, 18, 20, 25], rather than conducting the experiment around a table [20, 25] where only hand movements or upper body motions are collected, we implemented a free-standing setting in our setup. This configuration enables us to collect whole-body motion data. Indeed, a free-standing setting provides participants more freedom to perform body gestures to convey the verbal contents of their speech. Instead of creating interaction scenarios limited to a specific context, for instance, agree and disagree discussions [2], theatrical narratives [18], LISI-HHI covers a wider range of communication contexts. Importantly, LISI-HHI emphasises daily social HHI, which has many practical applications in HRI. For instance, available end-to-end learning frameworks, explained in Section 1, can be utilised to transfer human nonverbal communication skills to robots. Putting all together, LISI-HHI dataset aims to serve as a high accuracy and multimodal dataset for different research domains, especially social HRI.

3 THE LISI-HHI DATASET

3.1 Device Setup

The experiment was conducted in a motion capture room with devices set up as in Fig. 2. All sensors are synchronised together so they could be triggered and stopped simultaneously. The following describes the measured data and the sensors utilised in the setup:

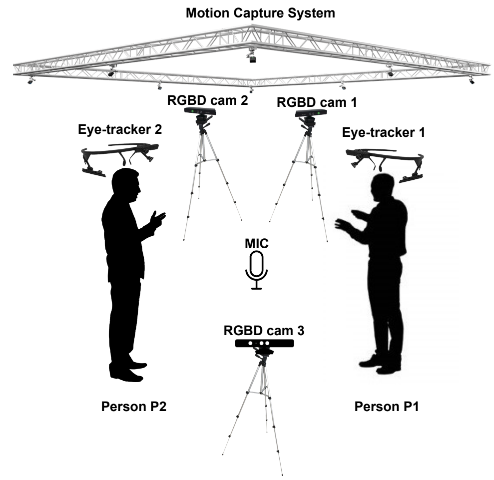


Figure 2: Data recording setup. We used 3 RGB-D cameras, 2 eye-trackers, 1 microphone, and a motion capture system with 8 Vicon infrared cameras.

- Motion capture: The system consists of 8 infrared cameras capturing the whole body motion at a frequency of 120 fps. Motion data $A^{39\times3\times T}$ of each participant consists of 39 joints presented in $3D$ coordinates, X, Y, Z, over the session time sequence T.
- Egocentric camera: The RGB camera attached to the participant's eye-tracker provides images at the frequency 30 fps with a resolution of 1080×1920.
- Eye-tracker: The eye gaze data is collected at a frequency of 30 fps. The data is presented by coordinate values aligned with RGB images of the corresponding egocentric camera.
- Static RGB-D camera: 2 RGB-D cameras are implemented to capture the individual semi-frontal field of view of each interlocutor. The third RGB-D camera is used for capturing the global side view. Images are captured at a resolution of 1080×1920 (30 fps).
- Audio: An omnidirectional microphone is placed in the middle of two participants to record their speech.

3.2 Interaction Scenarios

The dataset covers 32 dyads, and each dyad consists of 5 interaction sessions. With the aim of collecting a diverse set of verbal and nonverbal behaviours in different interaction contexts, participants are not given any narrations, and no constraints are put into them regarding their way of speaking and acting. The following briefly summarises the designed scenarios:

- **Small talk**: Two participants are instructed to start the conversation by introducing themselves to the other, followed by a random chit-chat (hobbies, weather, etc.). This session serves as an ice-breaker where participants can freely talk, discuss and get familiar with their interacting partners. The data collected from this task allow us to understand how a social conversation is initialised and how hand gestures (e.g., beat gestures [17]) are used in random chit-chat.
- **Meal planning**: Two interlocutors are required to discuss various options and then finalise a 5-course menu for their dinner,

Table 1: Summary of publicly available multimodal datasets of dyadic interaction.

Dataset	Interaction Setting	Participants	Interaction Sessions	Duration	Modality
MAHNOB Mimicry [1]	HHI: 2	40	54	11 hours	Audio, video, depth
USC-CreativeIT [18]	HHI:2	16	8	8 hours	Audio, video, motion data
MULAI [11]	HHI:2	26	13	5.9 hours	Audio, video, motion data, physiological signals
Talking with Hands 16.2M [16]	HHI:2	50	50	50 hours	Audio, motion data
JESTKOD [2]	HHI:2	10	98	4.3 hours	Audio, video, motion data
UDIVA [23]	HHI:2	147	188 × 5	90.5 hours	Audio, video, heart rate
M-MS [4]	HHI: 2	21	41 + 22 (only ECG)	16.2 hours	Video, audio, ECG
CMU Panoptic [13]	HHI: up to 8	-	-	5.5 hours	Video, audio, depth, motion data
MATRICS [20]	HHI: 4	40	10 × 3	9.2 hours	Video, audio, depth, motion data, eye-tracker, head accelerator
Rakovic et. al [8]	HHI: 2	6	6 × 4	-	Video, motion data, eye-tracker
DAMI-P2C [6]	HHI: 2	68	65	21.6 hours	Video, audio
MSP-IMPROV [3]	HHI: 2	12	6	9 hours	Video, audio
MIT Interview [19]	HHI: 2	69	138	10.5 hours	Video, audio
LISI-HHI	HHI: 2	64	32×5	8.3 hours	Video, audio, depth, motion data, eye-tracker

Table 2: Comparisons of available datasets, in term of modalities (SC: Static Camera, EC: Egocentric Camera, LC: Local Camera field of view, GB: Global Camera field of view, A: Audio, MC: Motion Capture, EG: Eye Gaze, ECG: Electrocardiographic signals)

Dataset	SC		EC	A	MC	EG	ECG
	LC	GB					
MAHNOB Mimicry[1]	✓	✓	X	✓	X	X	X
USC-Creative[18]	X	✓	X	✓	✓	X	X
MULAI[11]	✓	X	X	✓	X	X	✓
Talking with Hands 16.2M[16]	X	X	X	✓	✓	X	X
JESTKOD[2]	X	✓	X	✓	✓	X	X
UDIVA[23]	✓	✓	X	✓	X	X	X
M-MS[4]	X	✓	X	✓	X	X	✓
CMU Panoptic[13]	X	✓	X	✓	X	X	X
MATRICS[20]	✓	✓	✓	✓	✓	✓	X
Rakovic et. al[8]	X	X	✓	X	✓	✓	X
DAMI-P2C[6]	✓	✓	X	✓	X	X	X
MSP-IMPROV[3]	X	✓	X	✓	X	X	X
MIT Interview[19]	✓	X	X	✓	X	X	X
LISI-HHI	✓	✓	✓	✓	✓	✓	X

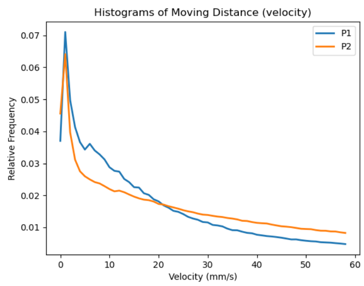
taking into consideration their interests and allergies. The scenario provides information about agreement and controversial behaviours towards completing a 5-course dinner menu [7].

- **Tangram game**: The game involves a director, who describes the game cards through their nonverbal gestures, and a follower, who will predict the cards the director explains. The session encourages the director and the follower to perform collaborative tasks for reproducing the orders of tangram cards. Indeed, the scenario allows us to collect a wide range of iconic and metaphoric gestures [17] that participants perform to describe the shape of tangram cards [10].
- **Role playing**: A customer is looking for a product at a shop. They ask a seller questions regarding the product specifications, warranty, etc., followed by negotiating a lower price for that item. This scenario provides a way to collect joint attention behaviours displayed by two interlocutors in a negotiation context.
- **Way finding**: A guest would like to know how to get to a coffee shop in the area which is familiar to the host. The host is asked to give detailed directions on how the guest can reach that coffee shop. This scenario aims to collect pointing and gaze gestures that the participants perform to navigate the area.

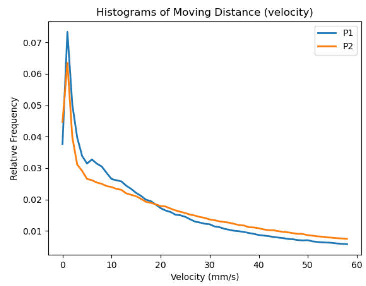
3.3 Data Collection Protocols

The experiment was conducted at a university and was approved by the King's College London Research Ethics Committee. We advertised the study through both internal and external call-for-participation websites. 170 participants, who showed interest in this study, were provided the information sheet with a full explanation of research objectives, experiment procedures, the anonymity of data collected, etc. They were asked to complete a pre-study questionnaire about their demographic information. After examining 170 questionnaires, we contacted 62 qualified participants to get their consent and schedule the experiment. We paired the participants in such a way that two interlocutors were in the same age category and in the same gender. It is because previous research [14, 26] suggests that different genders might introduce variability in behaviours in interpersonal coordination.

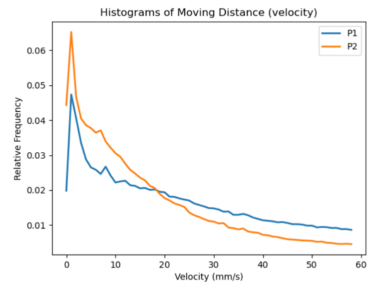
For conducting the experiment, 39 optical markers and an eye-tracker were attached to each participant. It was followed by motion capture and eye gaze calibration process. Participants were asked to



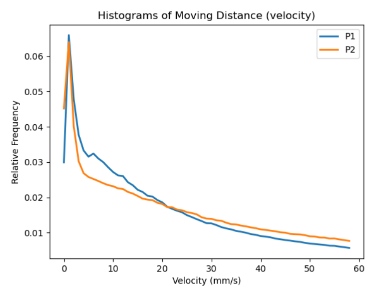
(a) Small talk



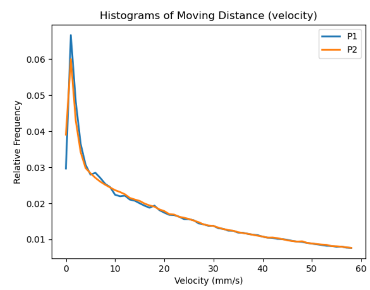
(b) Meal planning



(c) Tangram game



(d) Role playing



(e) Way finding

Figure 3: The average velocity distributions of participant 1, who played role as P1 (coloured in blue) and participant 2 assigned as P2 (coloured in orange). The velocity distributions were calculated based on the motion data collected from 64 participants.

conduct 5 interaction scenarios sequentially. Each interaction session was automatically triggered on and off in a fixed time interval. Experimenters were not allowed to be involved in the interaction sessions between two participants by any means. The experiment lasted for around an hour, including the experiment setting time.

3.4 Data Statistics

The dataset is composed of 8.3 hours of recording of dyadic interactions. 64 participants (38 females, 26 males) were assigned into 31 same-gender groups, each group was asked to conduct 5 interaction sessions. The majority of the participants (62%) were in the age group 18 to 24. Regarding cultural background, 46% participants identified themselves as Western people, while 41% participants reported that they have Asian culture. Most participants were students and staff members at universities in the UK, and 59% of them hold a bachelor's degree or higher.

Fig. 3 presents a summary of the motion statistics in the LISI-HHI dataset. We first calculated the average velocity of the whole body movements and then calculated the frequency of velocity values over time. Fig. 3 suggests that in all interaction sessions, except for *Tangram game*, the motion characteristics of person *P*1 and *P*2 are almost similar to each other. That could be explained by taking into consideration the interaction scenarios they involved. In the four sessions (*Small talk*, *Meal planning*, *Role playing*, and *Way finding*), the way *P*1 and *P*2 use nonverbal gestures to support their communication are almost similar. Vice versa, in session *Tangram game*, *P*2 is assigned as a director, who describes the game cards, while *P*1 plays a role as a follower, who predicts the card. Although body gestures are performed by both *P*1 and *P*2, *P*2 tends to act as many iconic gestures as possible in a limited time interval to describe the shape of the game card they are holding. Thus, the motion characteristic of *P*2 seems to be more extensive as compare to *P*1, who predicts the card based on the body gestures of *P*2.

4 CONCLUSION AND FUTURE WORK

This paper introduces LISI-HHI, a multimodal dataset of dyadic social interaction. The dataset consists of multiple nonverbal channels captured simultaneously from high-accuracy sensors. LISI-HHI is among a few databases recorded in English that cover multiple modalities, camera views, participants, and interaction sessions. The LISI-HHI dataset complements the previous databases by incorporating a multi-sensory setup with a novel design of multiple social interaction scenarios. We envision that LISI-HHI will contribute to the community as a reliable multimodal dataset that can be beneficial in various research areas, especially multimodal learning and social HRI. In future work, we will investigate the use of LISI-HHI to understand the dynamic exchange of social signals among interlocutors during the interaction. We will also consider a learning framework to transfer human nonverbal communication skills modelled from the LISI-HHI dataset into social robots.

ACKNOWLEDGMENT

This work was supported by the EPSRC project LISI (Grant Ref.: EP/V010875/1). The authors thank Andreea Barbinta, Chuyue Ding, and Narges Rahmani for their help with data collection and post-processing.

REFERENCES

[1] Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, and Maja Pantic. 2015. The MAHNOB Mimicry Database: A database of naturalistic human interactions. *Pattern recognition letters* 66 (2015), 52–61.

[2] Elif Bozkurt, Hossein Khaki, Sinan Keçeci, B Berker Türker, Yücel Yemez, and Engin Erzin. 2017. The JESTKOD database: an affective multimodal database of dyadic interactions. *Language Resources and Evaluation* 51, 3 (2017), 857–872.

[3] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing* 8, 1 (2016), 67–80.

[4] Gabriele Calabrò, Andrea Bizzego, Stefano Cainelli, Cesare Furlanello, and Paola Venuti. 2021. M-MS: A Multi-Modal Synchrony Dataset to Explore Dyadic Interaction in ASD. In *Progresses in Artificial Intelligence and Neural Systems*. Springer, 543–553.

[5] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. 2017. Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing* 10, 4 (2017), 484–497.

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

[7] Nicole Chovil. 1991. Discourse-oriented facial displays in conversation. *Research on Language & Social Interaction* 25, 1-4 (1991), 163–194.

[8] Nuno Ferreira Duarte, Mirko Rakovic, Jorge S Marques, José Santos-Victor, L Leal-Taixe, and S Roth. 2018. Action Alignment from Gaze Cues in Human-Human and Human-Robot Interaction.. In *ECCV Workshops (3)*. 197–212.

[9] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3497–3506.

[10] Judith Holler and Katie Wilkin. 2011. Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior* 35, 2 (2011), 133–153.

[11] Michel-Pierre Jansen, Khiet P Truong, Dirk KJ Heylen, and Deniece S Nazareth. 2020. Introducing MULAI: A multimodal database of laughter during dyadic interactions. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 4333–4342.

[12] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. 2020. Let's Face It: Probabilistic Multi-modal Interlocutor-aware Generation of Facial Gestures in Dyadic Settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[13] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2015. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*. 3334–3342.

[14] Johan C Karremans and Thijs Verwijmeren. 2008. Mimicking attractive opposite-sex others: The role of romantic relationship status. *Personality and Social Psychology Bulletin* 34, 7 (2008), 939–950.

[15] Jessica L Lakin, Valerie E Jefferis, Clara Michelle Cheng, and Tanya L Chartrand. 2003. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behavior* 27, 3 (2003), 145–162.

[16] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. 2019. Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 763–772.

[17] David McNeill. 2011. *Hand and mind*. De Gruyter Mouton.

[18] Angeliki Metallinou, Zhaojun Yang, Chi-chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan. 2016. The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations. *Language resources and evaluation* 50, 3 (2016), 497–521.

[19] Iftekhar Naim, M Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2015. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, Vol. 1. IEEE, 1–6.

[20] Fumio Nihei, Yukiko I Nakano, Yuki Hayashi, Hung-Hsuan Hung, and Shogo Okada. 2014. Predicting influential statements in group discussions using speech and head motion information. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 136–143.

[21] Lior Noy, Erez Dekel, and Uri Alon. 2011. The mirror game as a paradigm for studying the dynamics of two people improvising motion together. *Proceedings of the National Academy of Sciences* 108, 52 (2011), 20947–20952.

[22] Shogo Okada, Oya Aran, and Daniel Gatica-Perez. 2015. Personality trait classification via co-occurrent multiparty multimodal event discovery. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 15–22.

[23] Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio Junior, CS Jacques, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, et al. 2021. Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1–12.

[24] R Gnana Praveen, Eric Granger, and Patrick Cardinal. 2021. Cross attentional audio-visual fusion for dimensional emotion recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 1–8.

[25] Mirko Raković, Nuno Duarte, Jovica Tasevski, José Santos-Victor, and Branislav Borovac. 2018. A dataset of head and eye gaze during dyadic interaction task for modeling robot gaze behavior. In *MATEC Web of Conferences*, Vol. 161. EDP Sciences, 03002.

[26] Kikue Sakaguchi, Gudberg K Jonsson, and Toshikazu Hasegawa. 2005. Initial interpersonal attraction between mixed-sex dyad and movement synchrony. *The hidden structure of interaction: from neurons to culture patterns. Amsterdam* (2005).

[27] Nguyen Tan Viet Tuyen and Oya Celiktutan. 2021. Forecasting nonverbal social signals during dyadic interactions with generative adversarial neural networks. *arXiv preprint arXiv:2110.09378* (2021).

[28] Nguyen Tan Viet Tuyen and Oya Celiktutan. 2022. Agree or Disagree? Generating Body Gestures from Affective Contextual Cues during Dyadic Interactions. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1542–1547.

[29] Nguyen Tan Viet Tuyen, Armagan Elibol, and Nak Young Chong. 2020. Learning from humans to generate communicative gestures for social robots. In *2020 17th International Conference on Ubiquitous Robots (UR)*. IEEE, 284–289.