MultiMediate '23: Engagement Estimation and Bodily Behaviour Recognition in Social Interactions

Philipp Müller
DFKI
Saarbrücken, Germany
philipp.mueller@dfki.de

Michal Balazia*
INRIA Sophia Antipolis
Sophia Antipolis, France
michal.balazia@inria.fr

Tobias Baur*
University of Augsburg
Augsburg, Germany
tobias.baur@uni-a.de

Michael Dietz*
University of Augsburg
Augsburg, Germany
michael.dietz@uni-a.de

Alexander Heimerl*
University of Augsburg
Augsburg, Germany
alexander.heimerl@uni-a.de

Dominik Schiller*
University of Augsburg
Augsburg, Germany
dominik.schiller@uni-a.de

Mohammed Guermal
INRIA Sophia Antipolis
Sophia Antipolis, France
mohammed.guermal@inria.fr

Dominike Thomas
University of Stuttgart
Stuttgart, Germany
dominike.thomas@gmail.com

François Brémond
INRIA Sophia Antipolis
Sophia Antipolis, France
francois.bremond@inria.fr

Jan Alexandersson
DFKI
Saarbrücken, Germany
janal@dfki.de

Elisabeth André
University of Augsburg
Augsburg, Germany
elisabeth.andre@uni-a.de

Andreas Bulling
University of Stuttgart
Stuttgart, Germany
andreas.bulling@vis.uni-stuttgart.de

ABSTRACT

Automatic analysis of human behaviour is a fundamental prerequisite for the creation of machines that can effectively interact with- and support humans in social interactions. In MultiMediate '23, we address two key human social behaviour analysis tasks for the first time in a controlled challenge: engagement estimation and bodily behaviour recognition in social interactions. This paper describes the MultiMediate '23 challenge and presents novel sets of annotations for both tasks. For engagement estimation we collected novel annotations on the NOvice eXpert Interaction (NOXI) database. For bodily behaviour recognition, we annotated test recordings of the MPIIGroupInteraction corpus with the BBSI annotation scheme. In addition, we present baseline results for both challenge tasks.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**.

KEYWORDS

challenge, dataset, engagement, nonverbal behaviour

*These authors contributed equally to this work.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3613851

ACM Reference Format:
Philipp Müller, Michal Balazia*, Tobias Baur*, Michael Dietz*, Alexander Heimerl*, Dominik Schiller*, Mohammed Guermal, Dominike Thomas, François Brémond, Jan Alexandersson, Elisabeth André, and Andreas Bulling. 2023. MultiMediate '23: Engagement Estimation and Bodily Behaviour Recognition in Social Interactions. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3581783.3613851

1 INTRODUCTION

Artificial mediators [45], i.e. interactive intelligent agents that actively engage in a conversation in a human-like way have the potential to positively influence the course and/or outcomes of human interactions. They have been studied in a variety of contexts, including collaborative teamwork [10, 54], mental health [8], and education [16, 29]. A central prerequisite for effective and context-aware artificial mediation is the ability to comprehensively detect- and interpret the diverse set of social signals expressed by humans. At present, this challenge is still largely unsolved, and research on artificial mediators often has to rely on Wizard-of-Oz paradigms [8, 16, 29, 42, 51, 56].

With the multi-year MultiMediate challenge we contribute to realising the vision of autonomous artificial mediators by facilitating measurable advances on central conversational behaviour sensing and analysis tasks. The first iteration of the challenge in 2021 [37] has addressed eye contact detection and next speaker prediction while MultiMediate '22 has focused on backchannel analysis [1, 36]. In two separate tracks, MultiMediate '23 addresses the recognition of complex bodily behaviours, as well as the estimation of a persons' engagement level. Bodily behaviours such as fumbling, folded arms, or gesturing are a key social signal and were shown to be connected to many important high-level phenomena including stress regulation, attraction, or social verticality [13, 22, 33, 57]. As a result, accurate recognition of bodily behaviours can serve as a building block for the recognition of such more abstract phenomena. Knowing how engaged participants are, individually or as a group, is important for a mediator whose goal it is to keep engagement at a high level. Engagement is closely linked to the previous MultiMediate tasks of eye contact detection [41, 46] as well as backchanneling [20].

With MultiMediate '23 we present the first challenge on engagement estimation and the recognition of bodily behaviours in social interaction. We define the tasks and evaluation criteria and describe new annotations collected on the NOvice eXpert Interaction (NOXI) database [11], as well as on unreleased test recordings of MPIIGroupInteraction [39]. Furthermore, we present baseline approaches for both challenge tasks and report evaluation results. We make all collected annotations, baseline implementations, and raw feature representations publicly available for further use, even beyond the scope of MultiMediate '23.[1]

2 RELATED WORK

We review previous works on methods and datasets for engagement estimation and bodily behaviour recognition in social interaction.

2.1 Engagement Estimation

Engagement has been investigated from various research angles, e.g. how to define, annotate, or to automatically predict it. Rich et al. [48] introduced a module for the recognition of engagement in human-robot interaction based on backchannels. Sanghvi et al. [50] predicted engagement based on body posture features. Bednarik et al. [6] focused on recognizing conversational engagement with gaze data. Research in detecting engagement in students is prolific and promising [19, 25]. Engagement is also often studied in children [47] and, more particularly, in children interacting with an artificial agent [24, 40, 44]. Guhan et al. [21] researched engagement in mental health patients, based on videos of the patient. Some datasets also offer engagement ratings, such as RECOLA [49], MHHRI [14], and [23] with annotations from [6]. In Table 1 we provide an overview over the existing social interaction datasets with engagement annotations. The NoXi dataset annotated for MultiMediate '23 is significantly larger compared to previous datasets.

2.2 Bodily Behaviour Recognition

Bodily behaviours are key signals in social interactions and are related to many higher-level attributes. For example, displacement behaviours (e.g. fumbling, face-touching, or grooming) are associated with anxiety and stress regulation [5, 33, 34]. Leaning towards the interlocutor is connected with rapport [53] and crossed arms can be indicative of emotion expressions [60]. Further connections were found between bodily behaviours and liking [31, 32], attractiveness [57], and social verticality [22].

Despite this importance, little previous work addressed the recognition of bodily behaviours like fumbling, grooming, crossed arms, or gesturing in social interactions [3, 27]. While impressive progress was made on body- and hand pose estimation [12, 55], it is not a trivial task to establish the connection between low-level keypoint detections and complex bodily behaviours that are relevant to the interaction. Furthermore, only a limited number of bodily behaviour recognition datasets containing spontaneous behaviour in social interactions is available. The PAVIS Face-Touching dataset [7] consists of a single annotated behaviour (face touching) in group discussions. The iMiGUE dataset [27] contains annotations of 32 behaviour classes annotated for speakers at sports press conferences. For the purpose of MultiMediate , the recently published BBSI dataset [3] is most relevant, which consists of 15 behaviour classes annotated for all participants of 3-4 person group conversations. Such group conversations are one of the main application domains of artificial mediators.

[1]https://multimediate-challenge.org

Corpus	Screen	Group size	Length	Part.
Guhan et al. [21]	✓	2	1h5m	13
RECOLA [49]	✓	2	3h50m	46
Bednarik et al. [6]	✓	4-7	6h	9 groups
MMHRI [14]	✗	2	6h	18
NOXI (ours)	✓	2	25h	87

Table 1: Social interaction datasets with engagement annotations, excluding MOOC and school settings and children as participants. *Screen* indicates whether interaction was screen-mediated, *Group size* the number of humans per interaction, *Length* the total duration of interactions, and *Part.* the total number of human participants.

3 CHALLENGE DESCRIPTION

In the following we present the two challenge tasks and the utilised datasets. For both tasks test samples (without ground truth) are released to participants before the challenge deadline. Participants in turn submit their predictions for evaluation.

3.1 Engagement Estimation Task

Task definition. The task includes the continuous, frame-wise prediction of the level of conversational engagement of each participant on a continuous scale from 0 (lowest) to 1 (highest). Participants are encouraged to investigate multimodal as well as reciprocal behaviour of both interlocutors in the Novice-Expert Interaction corpus. We make use of the Concordance Correlation Coefficient (CCC) [26] to evaluate predictions on the test set.

Dataset. The NOvice eXpert Interaction (NOXI) database [11] is a corpus of dyadic, screen-mediated face-to-face interactions in an expert-novice knowledge sharing context. In a session, one participant assumes the role of an expert and the other participant the role of a novice. Figure 1 shows two users during interaction. NOXI includes interactions recorded at three locations (France, Germany and UK), spoken in eight languages (English, French, German, Spanish, Indonesian, Arabic, Dutch and Italian), discussing a wide range of topics. The dataset offers over 25 hours (x2) of recordings of dyadic interactions in natural settings, featuring synchronized audio, video (25fps), and motion capture data (using a Kinect 2.0).

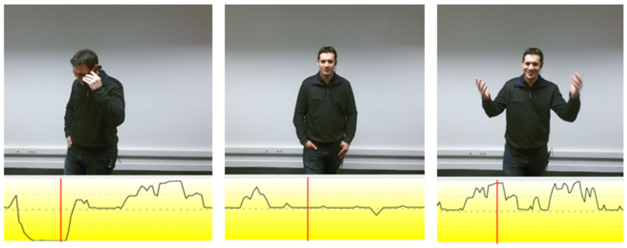

Figure 1: Snapshots of scenes of a participant in the NOXI corpus being disengaged (left), neutral (center) and highly engaged (right).

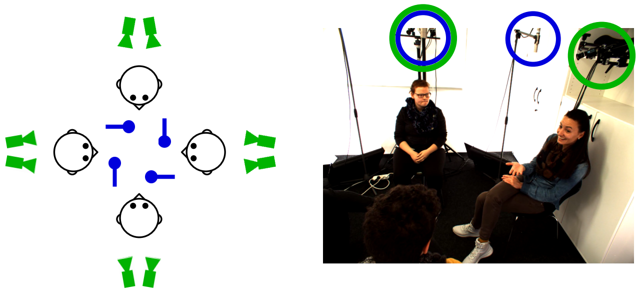

Figure 2: Setup of the MPIIGroupInteraction dataset. Reproduced with permission from the authors of [39].

We will use subset of this corpus containing 48 sessions for training and 16 sessions for testing (75/25 split). We aimed to obtain data of spontaneous behavior in a natural setting on a variety of discussion topics. Therefore, one of the main design goals was to match recorded participants based on their common interests. This means that we first gathered potential experts willing to share their knowledge about one or more topics they were knowledgeable and passionate about, and secondly we recruited novices willing to discuss or learn more about the available set of topics offered by experts. The corpus further introduces interruptions of the novices in order to provoke experts' reactions when conversational engagement gets interrupted. In particular, for this challenge, each session has been annotated in a continuous matter, meaning each video frame has a score between 0 and 1. Each rating was performed by at least two (up to 7) annotators (Average: 3.6 raters per session). We created gold standard annotations by calculating the mean over all raters. The NOXI dataset can be obtained from the website[2].

3.2 Bodily Behaviour Recognition Task

Task definition. We formulate bodily behaviour recognition as a multi-label classification task. Challenge participants are required to predict which of 15 behaviour classes are present in a 64 (2.13 sec) frame input window. For each 64-frame window, we provide a frontal view on the target participant, as well as two side views (left and right). As the behaviour classes on this task are highly unbalanced, we will measure performance using average precision computed per class and aggregated using macro averaging, i.e. giving the same weight to each class. This encourages challenge competitors to develop novel methods to improve performance on challenging low-frequency classes.

Dataset. As in MultiMediate '21 [37], our challenge is based on the MPIIGroupInteraction dataset [38, 39]. This dataset has served as a basis for diverse tasks, including emergent leadership detection [35], eye contact detection [18, 30, 38], next speaker prediction [9], backchannel analysis [1, 52], and body language detection [3]. The MPIIGroupInteraction corpus consists of 22 group discussions between three to four people, each lasting for 20 minutes [39]. This year's bodily behaviour task is based on the recently collected BBSI annotations [3], consisting of 15 bodily behaviour classes annotated on the whole MPIIGroupInteraction corpus. For MultiMediate '23, we excluded "Lean towards" as inter-annotator agreement was reported to be very low on this class. We collected bodily behaviour annotations for the remaining 14 classes on 996 samples obtained from six unpublished test recordings of MPIIGroupInteraction following the BBSI protocol [3]. To reach high-quality annotations on the test set, we obtained consensus decisions from three annotators. All classes except the "Stretching" class were present on the test set. The MPIIGroupInteraction dataset can be obtained from the website[3].

4 EXPERIMENTS AND RESULTS

We are providing a baseline model for each task. This section describes the training methodology as well as the utilized features and results achieved for both tasks.

4.1 Engagement Estimation

4.1.1 Approach. For the engagement estimation task we rely on a set of multimodal features comprising body posture, facial features and vocal features, followed by a fully connected neural network with three hidden layers of size 112 each. To prevent overfitting we rely on a dropout layer after the second hidden layer with a dropout rate of 0.25. The network has been trained using the Adam optimizer and the mean squared error loss function. All hyperparameters have been optimized using the hyperband search algorithm of the KerasTuner framework [43].

Head Features. We extracted features from participants' head and face using OpenFace 2.0 [4]. All features where extracted for each video frame. The resulting feature vectors are consisting of 68 3D facial landmarks, 56 3D eye landmarks, presence and intensity of 18 action units as well as markers for detection success, detection certainty facial position and rotation. Furthermore, we also use 17 action units provided by the Microsoft Kinect sensor.

Pose Features. We extract body pose estimates using OpenPose [12] as well as the Microsoft Kinect sensor data, resulting in the estimation of 350 data points comprising information about the location of various joints as well as their rotation.

Voice Features. For the paralinguistic assessment of engagement we extracted two feature sets over a one-second sliding window with a stride of 40ms to match the frame rate of the video stream. The

[2]https://multimediate-challenge.org/datasets/Dataset_NoXi/

[3]https://multimediate-challenge.org/datasets/Dataset_MPII/

Features	Val CCC	Test CCC
Head		
openface	0.23	0.21
AUs	0.31	0.22
Body		
skeleton	0.47	0.43
openpose	0.53	0.43
Voice		
gemaps	0.58	0.55
soundnet	0.54	0.49
Multimodal		
feature fusion + pca	**0.71**	**0.59**

Table 2: Concordance correlation coefficient (CCC) of our baseline on engagement detection validation and test sets.

Approach	Val MAP	Test MAP
random baseline	0.0884	0.2355
w/o bkgd class, frontal view	0.3974	0.5315
w/o bkgd class, side view 1	0.3030	0.4341
w/o bkgd class, side view 2	0.3628	0.4893
w/o bkgd class, max of views	0.4087	0.5333
w/o bkgd class, mean of views	0.4084	0.5402
w/ bkgd class, frontal view	0.4051	0.5498
w/ bkgd class, side view 1	0.3096	0.4451
w/ bkgd class, side view 2	0.3686	0.4641
w/ bkgd class, max of views	0.4062	0.5443
w/ bkgd class, mean of views	**0.4099**	**0.5628**

Table 3: Validation and test results for the random baseline and different variants of the Video Swin Transformer.

first feature set is the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [17]. This set consists of 54 acoustic parameters that are commonly applied to tasks like depression, mood, and emotion recognition [58]. Secondly, we used pretrained version of Soundnet [2] to extract sound embeddings from the raw signal. Soundnet is a deep convolutional neural network that has already been shown to provide effective features for vocal social signal analysis [59].

In our baseline approach, we fused the feature vectors of all modalities into one feature vector. As a large number of features can lead to overfitting we applied a PCA, reducing the number of features to 83 principal components.

4.1.2 Results. The results are depicted in Table 2. Among the single modalities the vocal features are clearly outperforming the body and head features on the validation set as well as on the test test. However, the multimodal feature fusion shows that the combination of all features still outperforms just using vocal features substantially. The additional value added by head and body features indicates that the expression of engagement is not clearly bound to one modality but should be analyzed considering multiple modalities.

4.2 Bodily Behaviour Recognition

4.2.1 Approach. As our baseline solution, we chose the Video Swin Transformer [28], which produced recent state-of-the-art results in action recognition tasks. It operates on fixed inputs of length 32 frames and size of 224×224 pixels. Given the input videos of length 64 frames and of larger resolutions, we set the stride to 2, that is we took every second frame, and we resized the video accordingly. We assigned input clips with multiple corresponding behavior class labels and clips of different viewpoints are treated as independent samples during training. To the clips with no labels, we assigned a new behavior class called *Background*, and, instead of the 14, trained the model in a 15-class multi-label setup. To aggregate predictions across views at test time, we averaged the scores obtained from all three views. We used the Swin Base model that is pre-trained on ImageNet and Kinetics-400, and fine-tuned it on the MPIIGroupInteraction dataset for only one epoch with learning rate 10^{-3} and with AdamW optimizer. Our implementation uses the open-source toolbox MMaction2 [15] built on top of PyCharm.

4.2.2 Results. Results of multiple ablations are reported in Table 3. We evaluated our approach against ablations that operate on single views, against an aggregation strategy using the maximum across views, as well as against not using an additional background class during training. The best mean average precision (MAP) on both validation and test sets was achieved by averaging across views and training with a background class. While the inclusion of the background class only led to minor improvements, averaging across views yielded consistent improvements. The best single view was the frontal view, and side views resulted in a significant performance drop. All results clearly outperformed the random baseline. Results on the test set tend are systematically higher, likely as a result of the higher quality annotations, and the lack of the "Stretching" class on the test set which as a result is always evaluated with 1.

5 CONCLUSION

We introduced MultiMediate '23, the first challenge addressing engagement estimation and bodily behaviour recognition in social interactions in well-defined conditions. We presented publicly available datasets and evaluation protocols for both tasks, and evaluated baseline approaches. The evaluation server will remain accessible to researchers even beyond the MultiMediate challenge, contributing to continuing progress on both tasks.

ACKNOWLEDGMENTS

P. Müller and J. Alexandersson were funded by the German Ministry for Education and Research (BMBF), grant number 01IS20075 and by the European Union Horizon Europe programme, grant number 101078950. A. Bulling was funded by the European Research Council (ERC; grant agreement 801708). M. Balazia was funded by the French National Research Agency under the UCA^{JEDI} Investments into the Future, project number ANR-15-IDEX-01. The researchers from Augsburg University were partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project Panorama, grant number 442607480.

REFERENCES

[1] Ahmed Amer, Chirag Bhuvaneshwara, Gowtham K Addluri, Mohammed M Shaik, Vedant Bonde, and Philipp Müller. 2023. Backchannel Detection and Agreement Estimation from Video with Transformer Networks. *arXiv preprint arXiv:2306.01656* (2023).

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. SoundNet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*. 892–900.

[3] Michal Balazia, Philipp Müller, Ákos Levente Tánczos, August von Liechtenstein, and François Brémond. 2022. Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation. In *Proc. of the ACM International Conference on Multimedia*. 70–79. https://doi.org/10.1145/3503161.3548363

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *Proc. of the IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 59–66. https://doi.org/10.1109/FG.2018.00019

[5] M Bardi, T Koone, S Mewaldt, and K O'Connor. 2011. Behavioral and physiological correlates of stress related to examination performance in college chemistry students. *Stress* 14, 5 (2011), 557–566. https://doi.org/10.3109/10253890.2011.571322

[6] Roman Bednarik, Shahram Eivazi, and Michal Hradis. 2012. Gaze and Conversational Engagement in Multiparty Video Conversation: An Annotation Scheme and Classification of High and Low Levels of Engagement. In *Proc. of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*. https://doi.org/10.1145/2401836.2401846

[7] Cigdem Beyan, Matteo Bustreo, Muhammad Shahid, Gian Luca Bailo, Nicolo Carissimi, and Alessio Del Bue. 2020. Analysis of face-touching behavior in large scale social interaction dataset. In *Proc. of the ACM International Conference on Multimodal Interaction*. 24–32. https://doi.org/10.1145/3382507.3418876

[8] Chris Birmingham, Zijian Hu, Kartik Mahajan, Eli Reber, and Maja J. Mataric. 2020. Can I Trust You? A User Study of Robot Mediation of a Support Group. *arXiv preprint arXiv:2002.04671* (2020).

[9] Chris Birmingham, Kalin Stefanov, and Maja J Mataric. 2021. Group-Level Focus of Visual Attention for Improved Next Speaker Prediction. In *Proc. of the ACM International Conference on Multimedia*. 4838–4842. https://doi.org/10.1145/3474085.3479213

[10] Dan Bohus and Eric Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *Proc. of the ACM International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. 1–8. https://doi.org/10.1145/1891903.1891910

[11] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel F. Valstar. 2017. The NoXi Database: Multimodal Recordings of Mediated Novice-Expert Interactions. In *Proc. of the International Conference on Multimodal Interaction*. https://doi.org/10.1145/3136755.3136780

[12] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 7291–7299. https://doi.org/10.1109/CVPR.2017.143

[13] Dana R. Carney, Dana. 2005. Beliefs about the nonverbal expression of social power. *Journal of nonverbal behavior.* 29, 2 (2005).

[14] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. 2019. Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement. *IEEE Transactions on Affective Computing* 10, 4 (2019), 484–497. https://doi.org/10.1109/TAFFC.2017.2737019

[15] MMAction2 Contributors. 2020. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2.

[16] Olov Engwall and José Lopes. 2020. Interaction and collaboration in robot-assisted language learning for adults. *Computer Assisted Language Learning* (2020), 1273–1309. https://doi.org/10.1080/09588221.2020.1799821

[17] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, 2 (2015), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

[18] Eugene Yujun Fu and Michael W Ngai. 2021. Using Motion Histories for Eye Contact Detection in Multiperson Group Conversations. In *Proc. of the ACM International Conference on Multimedia*. 4873–4877. https://doi.org/10.1145/3474085.3479230

[19] Patricia Goldberg, Ömer Sümer, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, and Ulrich Trautwein. 2021. Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction. *Educational Psychology Review* 33, 1 (2021), 27–49. https://doi.org/10.1007/s10648-019-09514-z

[20] Mononito Goswami, Minkush Manuja, and Maitree Leekha. 2020. Towards social & engaging peer learning: Predicting backchanneling and disengagement in children. *arXiv preprint arXiv:2007.11346* (2020).

[21] Pooja Guhan, Naman Awasthi, Kristin Bussell, Dinesh Manocha, Gloria Reeves, Aniket Bera, et al. 2020. Developing an Effective and Automated Patient Engagement Estimator for Telehealth: A Machine Learning Approach. *arXiv preprint arXiv:2011.08690* (2020).

[22] Judith Hall, Erik Coats, and Lavonia LeBeau. 2005. Nonverbal Behavior and the Vertical Dimension of Social Relations: A Meta-Analysis. *Psychological bulletin* 131 (12 2005), 898–924. https://doi.org/10.1037/0033-2909.131.6.898

[23] Michal Hradis, Shahram Eivazi, and Roman Bednarik. 2012. Voice activity detection from gaze in video mediated communication. In *Proc. of the ACM Symposium on Eye Tracking Research and Applications*. 329–332. https://doi.org/10.1145/2168556.2168628

[24] Shomik Jain, Balasubramanian Thiagarajan, Zhonghao Shi, Caitlyn Clabaugh, and Maja J Matarić. 2020. Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders. *Science Robotics* 5, 39 (2020). https://doi.org/10.1126/scirobotics.aaz3791

[25] Shofiyati Nur Karimah and Shinobu Hasegawa. 2021. Automatic Engagement Recognition for Distance Learning Systems: A Literature Study of Engagement Datasets and Methods. In *Augmented Cognition (Lecture Notes in Computer Science)*. Springer International Publishing, Cham, 264–276. https://doi.org/10.1007/978-3-030-78114-9_19

[26] Lawrence I-Kuei Lin. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 45, 1 (1989), 255–268. https://doi.org/10.2307/2532051

[27] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. 2021. iMiGUE: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 10631–10642. https://doi.org/10.1109/CVPR46437.2021.01049

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proc. of the IEEE/CVF International Conference on Computer Vision*. 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

[29] José Lopes, Olov Engwall, and Gabriel Skantze. 2017. A first visit to the robot language café. In *Proc. of the ISCA Workshop on Speech and Language Technology in Education*. https://doi.org/10.21437/SLaTE.2017-2

[30] Fuyan Ma, Ziyu Ma, Bin Sun, and Shutao Li. 2022. TA-CNN: A Unified Network for Human Behavior Analysis in Multi-Person Conversations. In *Proc. of the ACM International Conference on Multimedia*. 7099–7103. https://doi.org/10.1145/3503161.3551587

[31] Albert Mehrabian. 1968. Relationship of attitude to seated posture, orientation, and distance. *Journal of personality and social psychology* 10, 1 (1968), 26. https://doi.org/10.1037/h0026384

[32] Albert Mehrabian and John T Friar. 1969. Encoding of attitude by a seated communicator via posture and position cues. *Journal of Consulting and Clinical Psychology* 33, 3 (1969), 330. https://doi.org/10.1037/h0027576

[33] Changiz Mohiyeddini, Stephanie Bauer, and Stuart Semple. 2013. Displacement behaviour is associated with reduced stress levels among men but not women. *PloS one* 8, 2 (2013), e56355. https://doi.org/10.1371/journal.pone.0056355

[34] Changiz Mohiyeddini, Stephanie Bauer, and Stuart Semple. 2015. Neuroticism and stress: The role of displacement behavior. *Anxiety, stress, & coping* 28, 4 (2015), 391–407. https://doi.org/10.1080/10615806.2014.1000878

[35] Philipp Müller and Andreas Bulling. 2019. Emergent Leadership Detection Across Datasets. In *Proc. of the ACM International Conference on Multimodal Interaction*. 274–278. https://doi.org/10.1145/3340555.3353721

[36] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Hali Lindsay, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2022. MultiMediate'22: Backchannel Detection and Agreement Estimation in Group Interactions. In *Proc. of the ACM International Conference on Multimedia*. 7109–7114. https://doi.org/10.1145/3503161.3551589

[37] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Guanhua Zhang, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2021. MultiMediate: Multi-modal Group Behaviour Analysis for Artificial Mediation. In *Proc. of the ACM International Conference on Multimedia*. 4878–4882. https://doi.org/10.1145/3474085.3479219

[38] Philipp Müller, Michael Xuelin Huang, Xucong Zhang, and Andreas Bulling. 2018. Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour. In *Proc. of the ACM Symposium on Eye Tracking Research & Applications*. 1–10. https://doi.org/10.1145/3204493.3204549

[39] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In *Proc. of the ACM International Conference on Intelligent User Interfaces*. Association for Computing Machinery, 153–164. https://doi.org/10.1145/3172944.3172969

[40] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in Human-Agent Interaction: An Overview. *Frontiers in Robotics and AI* 7 (2020). https://doi.org/10.3389/frobt.2020.00092

[41] Catharine Oertel and Giampiero Salvi. 2013. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *Proc. of the ACM International Conference on Multimodal Interaction*. 99–106. https://doi.org/10.1145/2522848.2522865

[42] N. Ohshima, R. Fujimori, H. Tokunaga, H. Kaneko, and N. Mukawa. 2017. Neut: Design and evaluation of speaker designation behaviors for communication support robot to encourage conversations. In *Proc. of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 1387–1393. https://doi.org/10.1109/ROMAN.2017.8172485
[43] Tom O'Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi, et al. 2019. KerasTuner. https://github.com/keras-team/keras-tuner.
[44] Hae Won Park, Ishaan Grover, Samuel Spaulding, Louis Gomez, and Cynthia Breazeal. 2019. A Model-Free Affective Reinforcement Learning Approach to Personalization of an Autonomous Social Robot Companion for Early Literacy Education. In *Proc. of the AAAI Conference on Artificial Intelligence*. 687–694. https://doi.org/10.1609/aaai.v33i01.3301687
[45] Sunjeong Park and Youn-kyung Lim. 2020. Investigating User Expectations on the Roles of Family-shared AI Speakers. In *Proc. of the ACM Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3313831.3376450
[46] Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi. 2005. Engagement Capabilities for ECAs. *Autonomous Agents and Multi-agent Systems - AAMAS* (2005).
[47] Shyam Sundar Rajagopalan, O.V. Ramana Murthy, Roland Goecke, and Agata Rozga. 2015. Play with me — Measuring a child's engagement in a social interaction. In *Proc. of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Vol. 1. https://doi.org/10.1109/FG.2015.7163129
[48] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L Sidner. 2010. Recognizing engagement in human-robot interaction. In *Proc. of the ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 375–382. https://doi.org/10.1109/HRI.2010.5453163
[49] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proc. of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. https://doi.org/10.1109/FG.2013.6553805
[50] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W. McOwan, and Ana Paiva. 2011. Automatic Analysis of Affective Postures and Body Motion to Detect Engagement with a Game Companion. In *Proc. of the ACM/IEEE International Conference on Human-robot Interaction*. 305–312.
[51] Sarah Sebo, Brett Stoll, Brian Scassellati, and Malte F Jung. 2020. Robots in groups and teams: a literature review. *Proc. of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–36. https://doi.org/10.1145/3415247
[52] Garima Sharma, Kalin Stefanov, Abhinav Dhall, and Jianfei Cai. 2022. Graph-based Group Modelling for Backchannel Detection. In *Proc. of the ACM International Conference on Multimedia*. 7190–7194. https://doi.org/10.1145/3503161.3551605
[53] Christopher F. Sharpley and Anastasia Sagris. 1995. When does counsellor forward lean influence client-perceived rapport? *British Journal of Guidance & Counselling* 23, 3 (1995), 387–394. https://doi.org/10.1080/03069889508253696
[54] Elaine Short and Maja J. Mataric. 2017. Robot moderation of a collaborative game: Towards socially assistive robotics in group interactions. In *Proc. of the IEEE International Symposium on Robot and Human Interactive Communication*. 385–390. https://doi.org/10.1109/ROMAN.2017.8172331
[55] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 1145–1153. https://doi.org/10.1109/CVPR.2017.494
[56] Dina Utami and Timothy Bickmore. 2019. Collaborative user responses in multiparty interaction with a couples counselor robot. In *Proc. of the ACM/IEEE International Conference on Human-Robot Interaction*. 294–303. https://doi.org/10.1109/HRI.2019.8673177
[57] Tanya Vacharkulksemsuk, Emily Reit, Poruz Khambatta, Paul W Eastwick, Eli J Finkel, and Dana R Carney. 2016. Dominant, open nonverbal displays are attractive at zero-acquaintance. *Proceedings of the National Academy of Sciences* 113, 15 (2016), 4009–4014. https://doi.org/10.1073/pnas.1508932113
[58] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proc. of the International Workshop on Audio/Visual Emotion Challenge*. 3–10. https://doi.org/10.1145/2988257.2988258
[59] Johannes Wagner, Dominik Schiller, Andreas Seiderer, and Elisabeth André. 2018. Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant?. In *Proc. Interspeech*, B. Yegnanarayana (Ed.). 147–151. https://doi.org/10.21437/Interspeech.2018-1238
[60] Harald G Wallbott. 1998. Bodily expression of emotion. *European journal of social psychology.* 28, 6 (1998).