ConAn: A Usable Tool for Multimodal Conversation Analysis

Anna Penzkofer[*]
TU Munich
Munich, Germany
annapenzkofer@pm.me

Philipp Müller
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
philipp.mueller@dfki.de

Felix Bühler
University of Stuttgart
Stuttgart, Germany
st117123@stud.uni-stuttgart.de

Sven Mayer
LMU Munich
Munich, Germany
info@sven-mayer.com

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

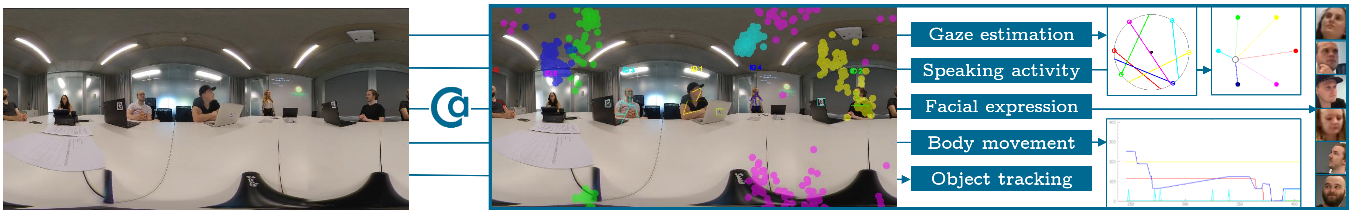


Figure 1: *ConAn* – our graphical tool for multimodal conversation analysis – takes 360 degree videos recorded during multi-person group interactions as input. *ConAn* integrates state-of-the-art models for gaze estimation, active speaker detection, facial action unit detection, and body movement detection and can output quantitative reports both at individual and group level, as well as different visualizations that provide qualitative insights into group interaction.

ABSTRACT

Multimodal analysis of group behavior is a key task in human-computer interaction, and in the social and behavioral sciences, but is often limited to more easily controllable laboratory settings or requires elaborate multi-sensor setups and time-consuming manual data annotation. We present *ConAn* – a usable tool to explore and automatically analyze non-verbal behavior of multiple persons during natural group conversations. In contrast to traditional multi-sensor setups, our tool only requires a single 360° camera and uses state-of-the-art computer vision methods to automatically extract behavioral indicators, such as gaze direction, facial expressions, and speaking activity. As such, our tool allows for easy and fast deployment and supports researchers in understanding individual behavior, group interaction dynamics, and in quantifying user-object interactions. We illustrate the benefits of *ConAn* on three sample use cases: conversation analysis, assessment of collaboration quality, and impact of technology on audience behavior. Taken together, *ConAn* represents an important step towards democratizing automatic conversation analysis in HCI and beyond.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

graphical user interface, conversation analysis, non-verbal behavior, group interaction

ACM Reference Format:
Anna Penzkofer, Philipp Müller, Felix Bühler, Sven Mayer, and Andreas Bulling. 2021. *ConAn*: A Usable Tool for Multimodal Conversation Analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3462244.3479886

1 INTRODUCTION

The need to sense, analyze, and understand conversations within groups of people arises in a range of different areas in HCI, computer-supported cooperative work (CSCW) as well as the social and behavioral sciences. In HCI, conversation analysis is important for applications such as conversational agents, human-robot interaction, or virtual and augmented reality [61, 77]. However, conversation analysis is still a largely manual and thus time-consuming, cumbersome, and error-prone process both to setup and to perform. For example, Chattopadhyay et al. [24] manually analyzed observations, interviews, and usage logs as well as video recordings to study group behavior while Brown et al. [20] studied phone conversations using manual labeling of more than 24h of video material. In addition, researchers have argued that manual coding of videos is likely to be influenced by the coder, resulting in annotation bias [70].

[*] Research conducted while at the University of Stuttgart

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479886

Modality	Usage
Eye Gaze	disruptive impact of mobile interactions [58], rapport detection [66], emergent leadership detection [18, 19, 63], social plane during interaction [73], mobile internet search as a part of conversation [20], interpersonal relationships at work [91], shared attention [61], classroom attention [75], human-robot-interaction [28], gaze cuing [71], social anxiety disorder [84], autism spectrum disorder [80], turn-taking [45], conversational engagement [16], social phobia [9], work group mood [12]
Speaking Activity	paralinguistic persuasion [94], emergent leadership detection [18, 63], rapport detection [66], personality trait prediction [51, 69], collective intelligence [32, 102], emotion recognition [62], prediction of extraversion trait [51], Meeting Mediator: feedback for group collaboration [49], interspeaker influence [23]
Body Pose and Hand Movement	social rank [100], attention detection [1], assessment of public speaking skills [25], in-class student participation [76], instructional communication [93], work group mood [12], behavior imitation [30], group emotional contagion [11], teaching behavior [46]
Facial Expression	social rank [100], rapport detection [66], classify human-human versus human-machine interaction [67], collective intelligence and group satisfaction [26], bias detection [50], group emotional contagion [11], autism spectrum disorder [80], behavior imitation [30]
Environment Tracking*	in the wild understanding [20], technology use in conversations [72, 89], device tracking [58], device logs [24]

Table 1: Common modalities in conversation analysis. * short list of factors mainly relevant for the HCI domain.

Analysis challenges due to the variability of human behavior are amplified in interactions between multiple people, and thus, the analysis of multiple modalities is required, such as speech, body language, and gaze. This has triggered research to move away from manual, time-consuming analysis, as well as the need for specialized hardware [16, 24] towards using computational methods for automatic analysis that only require off-the-shelf sensors [19, 82]. Over the last decade, various computational methods have been developed that enable researchers to automatically extract conversation characteristics from audio and video footage, e.g., rapport [66], leadership [63], and eye contact [8, 43, 64]. However, these methods have been developed independently and therefore, are largely inaccessible to the aforementioned research communities. Finally, the analysis of individual characteristics by using separate tools makes it challenging to jointly analyze and identify similarities, correlations, or patterns across modalities and people. Thus, the deployment of multiple sensors and the usage of multiple analysis tools is hard [40].

To address these limitations, and democratize the use of multimodal conversation analysis in HCI and beyond, we introduce *ConAn* – a usable tool that helps researchers in exploring and automatically analyzing non-verbal behavioral data of multiple people during natural group conversations. To support usability and foster easy deployment, our tool only requires a single 360° camera and integrates state-of-the-art algorithms for eye gaze estimation, action unit detection, speaker diarization, object tracking, and non-verbal gestures based on the body pose under a single graphical user interface. The development of *ConAn* was informed by a literature review to identify major features and dominant use cases of automatic conversation analysis across multiple disciplines. We illustrate the application of our tool by discussing multiple use cases it enables: general conversation analysis including observations on the disruptiveness of mobile phones, assessment of collaboration quality on a recorded collaboration task between three participants, and a small study on the impact of technology on audience behavior by comparing two videos with the same scenario, one with and the other without devices. We hereby showcase the use and flexibility for researchers within HCI but also how *ConAn* could be used for general social and behavioral research.

In summary, this work contributes an open-source tool that allows to easily study conversations with multiple people using a single 360° camera. We offer a wide range of state-of-the-art non-verbal and verbal analysis possibilities, such as gaze estimation and speaker diarization. Our tool is also equipped with the functionality to understand the impact of technologies surrounding conversation partners. This allows HCI researchers to rapidly analyze group scenarios without the need for elaborate camera and microphone setups. Moreover, the tool can serve as a screening tool for social psychologists. In conjunction with providing researchers with the open-source tool, we also provide 4 videos as a showcase evaluation highlighting the easy use of our setup and tool.

2 RELATED WORK

Conversation analysis is conducted in many domains, such as human resources development, education, or HCI. In the following, we first discuss the modalities that current analyses commonly use. Afterwards, we discuss how conversation analysis is used in HCI to better understand the interplay between humans and technology. Finally, we present currently available conversation analysis tools.

2.1 Modalities in Conversation Analysis

A number of fields use conversation analysis such as psychology [9, 84], human resources development [6, 53], learning analytics [73–75, 93], social skills training [13], or human robot interaction [28, 61]. Despite many similarities in terms of the general approach, we identified specific differences in the type of modalities that these fields primarily use, as well as the specific features that are extracted from them, see Table 1 for an overview.

Eye Gaze, usually in the form of eye contact, provides rich information on conversational behavior, such as engagement [16] or workgroup mood [12]. Eye gaze can also be used to infer the presence of personality traits [42], social anxiety disorder [84] or autism spectrum disorder [80]. Gaze further provides insights into how individuals behave as part of the group in terms of turn taking [45] or how pairs of people interact with one another, cf. [65].

Speaking Activity is widely studied during social interactions, including inter-speaker influence [23], paralinguistic persuasion [94], but also in the context of interaction mediation systems that try to improve collaboration [49]. Furthermore, active speech is closely correlated with non-verbal features like gestures [33] and eye contact [41, 48, 78]. These close connections indicate that social interaction understanding can benefit from jointly analyzing speaking activity with visual feature channels. While single-modal approaches only using audio [7, 35, 98] or video [36, 85] work well, multi-modal approaches are even more promising [10, 27, 37].

Body Pose and Hand Movement are important in conversations, e.g., for speakers to convey their message [15], in instructional communication [93], or for assessing of public speaking skills [25]. While body and hand gestures can repeat verbally stated arguments, they do not always need to [15]. Therefore, it is important to not only understand the verbal communication but also the accompanying non-verbal communication. For instance, deictic gestures are a core element of non-verbal expressions to guide the partners' attention [21, 38, 59]. Body and hand pose are also important for understanding classroom participation in educational settings [1, 44]. Various computer vision methods have been proposed to determine the human pose based on a RBG image [22, 90, 92].

Facial Expressions can both uncover general mood within a group but also information about individuals, such as the social rank [100] or low rapport detection [66]. Other researchers extracted group features like satisfaction [26] and emotional contagion [11]. Facial expressions were also shown to be usable for uncovering general personality disorders, e.g., the reaction times of facial expressions can detect potential threats to self or to close others [50]. Some approaches for extracting facial expressions make use of depth cameras to better support the recognition [52, 99] while other approaches can even extract the underlying action units such as [4, 54, 86].

2.2 Conversation Analysis in HCI

Wobbrock [101] provides an excellent review of how computer and technology in the environment can distract and impair people. Such impacts are for instance mobile phone use in group settings [89] or even in bed [81]. To understand these impacts, HCI research used a wide range of features based on conversation analysis. For instance, Mayer et al. [58] used gaze direction combined with video observation and interviews to understand the impact of incoming calls on a face-to-face conversation. Moreover, the quality of support induced by technology can be measured. For instance, Bednarik et al. [16] used only gaze to measure engagement in multi-group video calls and Chattopadhyay et al. [24] studied group behavior using observation, interviews, and usage logs in combination with video recordings. Brown et al. [20] analyzed mobile search in everyday conversations via video recordings of phone use and conversation transcripts between participants. Based on these insights, social disruption can be improved during design time [68].

2.3 Conversation Analysis Tools

In an attempt to reduce the complexity associated with recording and analyzing social behavior, several toolkits were proposed in a variety of scenarios [60, 83, 87]. We summarize the most relevant toolkits that apply to conversation analysis in Table 2.

The Social Signal Interpretation (SSI) framework [96] supports synchronized recording from multiple sensors as well as plug-in detection algorithms as a basis for behavior analysis. Complementary to SSI, NOVA [14, 39] focuses on the annotation process of multimodal behavior and provides functionalities for joint human-machine annotation. As NOVA is specifically built for the annotation use case, it does not provide visualizations or summary statistics for group behavior dynamics based on gaze behavior or speaking distribution which allow for a rapid, user-friendly understanding of recorded interactions. Based on the NOVA and SSI frameworks, more specialized tools answer the need for use case-specific functionalities [5, 88]. For instance, TARDIS [5] is targeted at (dyadic) job interview training in human-avatar interactions and offers playback of recordings obtained with webcam, kinect and microphone along with visualizations of annotations. On the other hand, MultiSense [88] specializes on the use case of psychological distress analysis in dyadic interactions, providing online- as well as offline feedback. Recently, Stefanov et al. [87] introduced OpenSense: a real-time multimodal acquisition and recognition system of social signals. OpenSense offers different components which can be loaded in a pipeline editor for selecting modalities of interest for each use case. Similar to NOVA, OpenSense supports analyzing multiple modalities with out-of-the-box visualizations, but needs separate pipelines for multiple subjects and therefore does not provide visualizations of group dynamics.

In summary, while several frameworks applicable to multimodal conversation analysis have been proposed, none of these frameworks provides a user-friendly, out-of-the-box solution for group interaction analysis. This is due to the lack of explicit support of a simplified sensor setup (e.g., a single 360° camera) as well as the ability to use case-specific analyses and visualizations (e.g., eye contact and speaking distribution). In contrast, *ConAn* provides an out-of-the-box solution for group behavior analysis with a simplified sensor setup and key analysis and visualization functions. At the same time, *ConAn* maintains flexibility via a modular design.

3 CHALLENGES IN MULTIMODAL GROUP INTERACTION ANALYSIS

The main challenge in conversation analysis is the lack of available datasets which in turn is due to the need for time-consuming data annotation and elaborate multi-sensor setups. For instance, Müller et al. [66] used eight cameras and four microphones, resulting in 12 separate recording devices which needed to be synchronized and oriented towards pre-defined seating positions of participants. Likewise, Beyan et al. [19] used four cameras and microphones, and an additional camera capturing the whole scene for data annotators. These controlled data collection approaches prevent research from being conducted in settings analogous to real-life situations.

Even though a 360° camera overcomes the aforementioned challenges, it also poses a new technical challenge: the lens distortion.

Name	Target Use Case	Modalities	Open Source	Multi-Platform	360° Support
MutualEyeContact [83]	Dyadic Interaction Analysis	Gaze, Facial Expressions	✗	Windows	✗
SSI [96]	Multimodal data recording and feature extraction	Extendable multi-sensor recording framework	✓	Windows*	✗
NOVA [14]	Annotation & cooperative machine learning	Extendable annotation framework	✓	Windows	✗
MultiSense [88]	Analysis of dyadic counseling interactions	Speech, Body, Gaze, Face	✗	Windows	✗
TARDIS [5]	Job interview training	Speech, Body, Gaxe, Face	✗	Windows	✗
OpenSense [87]	Multimodal data recording and feature extraction	Gaze, Speech, Body Pose, Head Gestures, Facial Expressions, Music	✓	Windows	✗
ConAn	Group Interaction Analysis	Gaze, Speaking Status, Facial Expressions, Body Pose, Object Tracking	✓	✓	✓

Table 2: Conversation Analysis Tools (* Linux & Mac via mobileSSI https://github.com/hcmlab/mobileSSI)

To be able to extract face crops as input to various models, a perspective transformation has to be performed with the center of each face as a reference point. Consequently, each subject's position needs to first be detected and tracked. To allow for people to move freely while still being able to determine their field of view (FoV) a geometric model of the room is necessary. In particular, the position of each subject needs to be set as the starting point for their gaze vector which then needs to be projected back onto the image plane. Without a geometric model each subject needs to remain in a fixed position, as e.g. in Müller et al. [64] for eye contact detection. Moreover, the design and usability of a graphical user interface (GUI) needs to be in line with the desired task to fulfill, while still being general enough for different scenarios. For multimodal conversation analysis, the main interest lies in the interaction between conversation partners. However, not many visualizations for interaction analysis on a group level have been proposed so far. On the other hand, many non-verbal feature extraction models exist, but each model requires a separate pre-processing pipeline and therefore also adds processing time.

Overall, the identified key challenges are the supported technical collection setup requiring complex processing steps, at present limiting group behavior recording and analysis to experts, as well as the conceptual design of a usable GUI with intuitively understandable visualizations and key modalities selected based on their importance for the target use case.

4 DESIGN OF *CONAN*

Our literature review revealed that conversation analysis is done using four key modalities to extract a multitude of insights into the individual conversation partners, the relationships between them, as well as overall insights into the conversation. However, conversation analysis is mostly a tedious and time-consuming task. At the same time, we discussed how current advances in computer vision and machine learning allow for automatic extraction of the modalities upon which these conversation insights are built.

In the following, we present *ConAn*: a tool that provides state-of-the-art machine learning models in an easy to use GUI, see Figure 1 and 2, enabling researchers to perform fast conversation analysis. Time is of the essence, especially during rapid prototyping and design sessions; both common tools in the HCI domain. Because our system requires only footage from a single 360° camera to capture all salient aspects of a conversation, users overcome the limitations of time-consuming annotation procedures.

Our system is designed to take every 360° video in an equirectangular projection and conversation audio as input. On this video and audio we then perform several pre-processing and extraction steps to ultimately provide them to the user in a GUI. Moreover, we developed *ConAn* using Qt[1] for cross-platform support. In the following, we discuss in detail which models and tools we used. However, our system structure is modular and allows for the replacement of individual models in the upcoming years to make use of the latest developments and advancements in machine learning and conversation analysis.

The source code for *ConAn* is available under MIT license via our git repository[2]. This allows other researchers to effectively and efficiently perform conversation analysis and thus, spark new investigations to improve the interplay between humans and technology.

4.1 User Interface

For usability we split the GUI into areas each with its own theme. On the left side of the upper half (see Figure 2 I) the video is displayed. We use video overlays to display labels for all participants, their body pose and gaze targets, as well as detected object locations. These overlays are visualized in Figure 3a, 4c, and 3e respectively, and can be toggled on or off at the control panel on the left side of the video (see Figure 2 II). A multi-segment selector is positioned below the video (see Figure 2 III). By default, the whole video is selected in one segment. By dragging the green and red separators, the start and end of the segment can be changed. Additional segments can be added by double-clicking on the empty region while removing a segment is possible by double-clicking on an existing segment. Below the multi-segment selector is a play/pause button and a standard video timeline slider.

The lower half, as shown in Figure 2 IV, consists of multiple tabs, with a separate tab for each of the five analyzed modalities. For each modality we display a set of aggregated features, as well as a dynamic visualization of the underlying data. If the selected

[1] https://www.qt.io/

[2] https://www.perceptualui.org/publications/penzkofer21_icmi/

segments are changed (via Figure 2 III), aggregated features are updated accordingly. Following [95], the data of all subjects is displayed as default, but the visualization of each subject can be hidden or shown again with the corresponding checkbox in Figure 2 II.

In the eye gaze tab (see Figure 3b) the yaw gaze of each subject and their position is visualized from a top-down view to enable the users of *ConAn* to capture eye contact behavior at first glance. Additionally, features calculated based on the amount of tracked frames, indicate various measures, including the relative time a subject is looked at by others, the time a subject spends looking at other people, the amount of time a subject is not looking at others, and a subject’s ratio between being watched and looking at other people. These features are commonly used for conversation analysis tasks, such as emergent leadership detection [63].

The distribution of speaking time is visualized dynamically, i.e. the total amount up to the current frame, with a moving circle indicating the balance of group conversation (see Figure 3d). This design was based on a real-time feedback application for enhancing group collaboration [49]. Speaking features commonly used in previous work [66] are extracted and displayed next to the graph. Features are composed of the total amount of time a subject is speaking in relation to the video length, the number of speaking turns, where one speaking turn is defined as the consecutive time a person is actively speaking, the average duration of all such speaking turns, and the average number of speaking turns per minute.

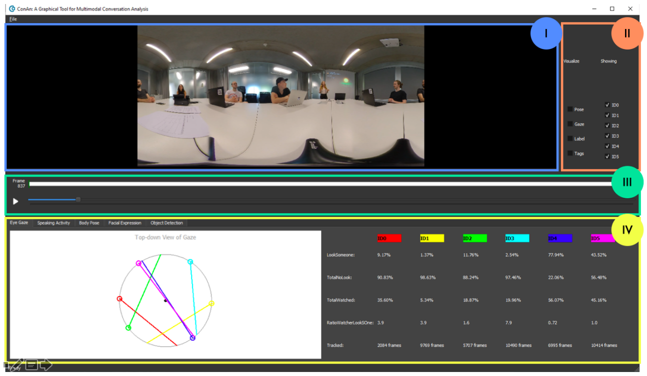
Figure 3f shows that the absolute body movement is displayed dynamically over time in terms of euclidean distance between first and current frame position, similar to [95]. Additionally, the frame number with the largest body activity is shown as a variable to enable users to quickly select interesting time segments. For hand movement, we selected three features, namely the relative time both hands were above the table, the relative time the movement of both hands exceeded a velocity threshold, and the hand velocity in the current frame, which is defined as the change of hand positions between current and last frame.

In the facial expression tab, similar to [4], cropped face images of each subject are displayed to highlight the detected facial expressions as coded by the Facial Action Unit Coding System (FACS)[3]. With our selected approach we are able to extract 12 different action units, for each of which the detected probabilities in the current frame are shown next to the image of the respective participant (see Figure 4b).

The movement of objects, similar to the visualization in the body movement tab, is displayed in a dynamic line graph (see Figure 4d). Each object has a unique tag id, which can be supplemented with available context information in an editable text field next to each tag. Additionally, the percentage of tracked frames versus overall video frames is shown on the right.

4.2 Gaze Estimation

For gaze estimation in-the-wild various options are available, including Gaze360 [47], OpenGaze [103], OpenFace [4], and RT-GENE [34][4]. While considering our application, Gaze360 [47] seems to fit best, the results were not satisfying. Thus, we finally decided to use RT-GENE [34]. In addition to feeding each video frame to the model, we also input a version of the frame where the left side and the right side are wrapped together. This enables us to detect when a person moves over the edge of the video, as none of the models account for this. Moreover, as this is a single frame estimation, we then track all subjects throughout the video using a minimal euclidean distance heuristic. Finally, to reduce outliers and noise due to the single frame estimation, we apply a rolling average window with a window length of $\frac{1}{3}$ of a second (10 frames at 30FPS).



	ID0	ID1	ID2	ID3	ID4	ID5
LookSomeone:	9.17%	1.37%	11.76%	2.54%	77.94%	43.52%
TotalNoLook:	90.83%	98.63%	88.24%	97.46%	22.06%	56.48%
TotalWatched:	35.60%	5.34%	18.87%	19.96%	56.07%	45.16%
RatioWatcherLookSOne:	3.9	3.9	1.6	7.9	0.72	1.0
Tracked:	2084 frames	9769 frames	5707 frames	10490 frames	6995 frames	10414 frames

Figure 2: System *ConAn*: layout of user interface.

From the gaze direction, we can then visualize information such as if the person is engaged in the conversion or if persons are looking at each other.

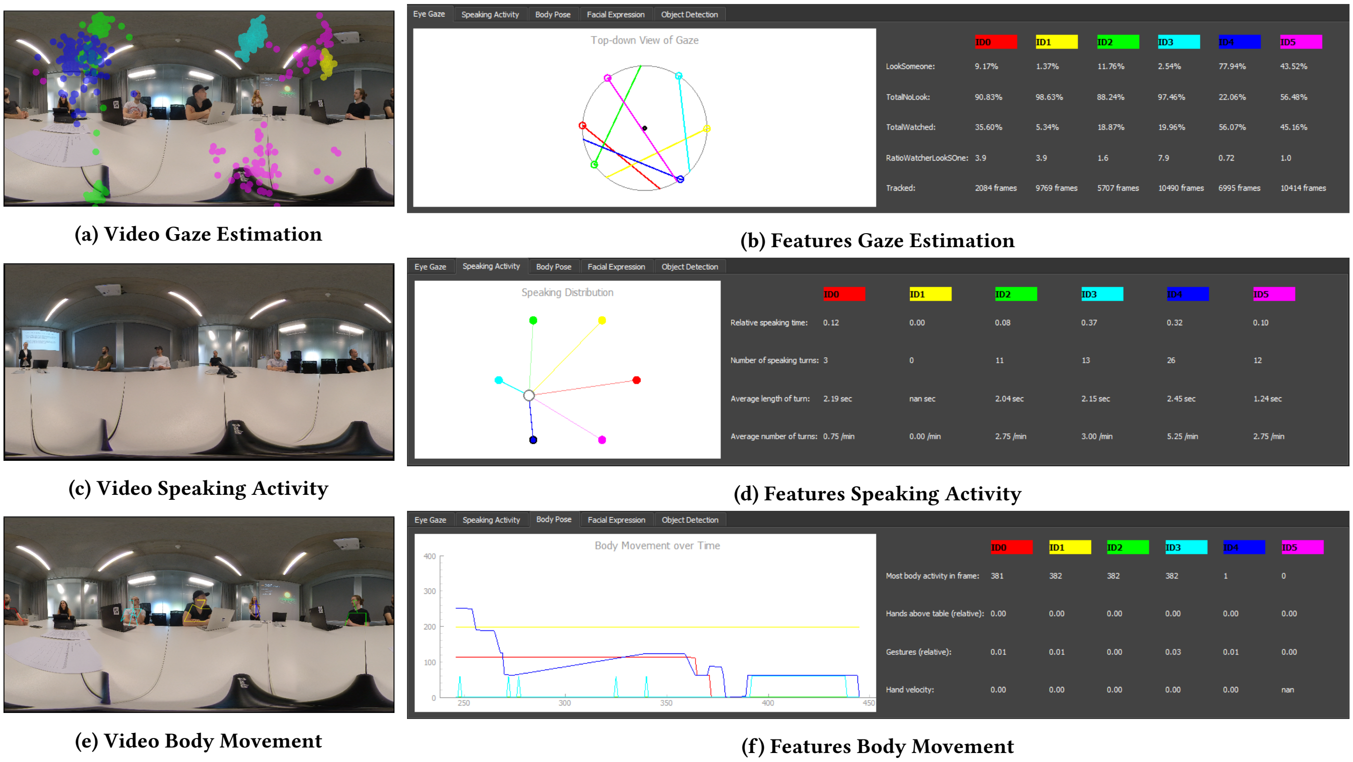
4.3 Body Movement

To be able to analyze body movement, a person’s body pose has to be detected first. While there has been a lot of research concerning this task, e.g., DeepPose [92], PandaNet [17], or OpenPose [22] proved to be the most easily accesible by providing a variety of pretrained models. OpenPose is a multi-person keypoint detector that is runtime invariant to the number of people in one frame. For our case, we used the 18-keypoint model, which takes the full frame as input and jointly predicts anatomical keypoints and a measurement for the degree of association between them. Based on the predicted degree of association keypoints are assigned to each person yielding a 2D skeleton, as can be seen in Figure 3e. Then, each identity is tracked throughout all frames of the video with a minimal euclidean distance heuristic. As the neck keypoint of each subject was the most consistently detected, its location was used to calculate overall body movement by taking its relative euclidean difference between frames. The location of both wrists is further used to track hand movement. Today’s 3D skeleton estimation models do not take fingers into account; however, when using OpenPose [22] the 2D joint estimation could be extended to detect fine grained hand poses such as reading a book, relaxed, and prayer [2].

4.4 Facial Action Unit Detection

Facial action units are based on an anatomical analysis of the face and can be described according to the (FACS) defined by Ekman et al. [31]. There are many different options available for detecting

[3] https://www.cs.cmu.edu/~face/facs.htm
[4] https://github.com/Tobias-Fischer/rt_gene/



(a) Video Gaze Estimation

	ID0	ID1	ID2	ID3	ID4	ID5
LookSomeone:	9.17%	1.37%	11.76%	2.54%	77.94%	43.52%
TotalNoLook:	90.83%	98.63%	88.24%	97.46%	22.06%	56.48%
TotalWatched:	35.60%	5.34%	18.87%	19.96%	56.07%	45.16%
RatioWatcherLookSOne:	3.9	3.9	1.6	7.9	0.72	1.0
Tracked:	2084 frames	9769 frames	5707 frames	10490 frames	6995 frames	10414 frames

(b) Features Gaze Estimation

(c) Video Speaking Activity

	ID0	ID1	ID2	ID3	ID4	ID5
Relative speaking time:	0.12	0.00	0.08	0.37	0.32	0.10
Number of speaking turns:	3	0	11	13	26	12
Average length of turn:	2.19 sec	nan sec	2.04 sec	2.15 sec	2.45 sec	1.24 sec
Average number of turns:	0.75 /min	0.00 /min	2.75 /min	3.00 /min	5.25 /min	2.75 /min

(d) Features Speaking Activity

(e) Video Body Movement

	ID0	ID1	ID2	ID3	ID4	ID5
Most body activity in frame:	381	382	382	382	1	0
Hands above table (relative):	0.00	0.00	0.00	0.00	0.00	0.00
Gestures (relative):	0.01	0.01	0.00	0.03	0.01	0.00
Hand velocity:	0.00	0.00	0.00	0.00	0.00	nan

(f) Features Body Movement

Figure 3: The in depth view of the gaze, speaking, and body features in the view (IV) of *ConAn*.

facial action units: OpenFace [4], AU R-CNN [54], or combining face alignment with action unit detection, as proposed in JAA-Net [86][5]. These approaches mainly differ in the amount of action units they are able to extract. Therefore, to maximize the number of extractable action units, we decided to use JAA-Net [86]. We trained JAA-Net on DISFA: a database of facial action units including intensities, made available by Mavadati et al. [56, 57]. After training, the model is able to perform face alignment by predicting facial landmarks, as well as global and local feature extraction for facial action units. The model takes cropped face images as input, which we extracted from the video frame for each subject based on the RT-GENE [34] head center prediction. The model then outputs activation maps, predicted landmarks, and predicted probabilities of 12 different action units: inner and outer brow raiser, brow lowerer, upper lid raiser, cheek raiser, nose wrinkler, lip corner puller, which is also known as a smile, lip corner depressor, chin raiser, lip stretcher, lips part, and jaw drop. We display these predicted probabilities as seen in Figure 4b. The action units are defined as present if the probability exceeds $\frac{1}{2}$ and the amount of frames in which a specific action unit is present for each person is included in the export file.

4.5 Speaker Activity Detection

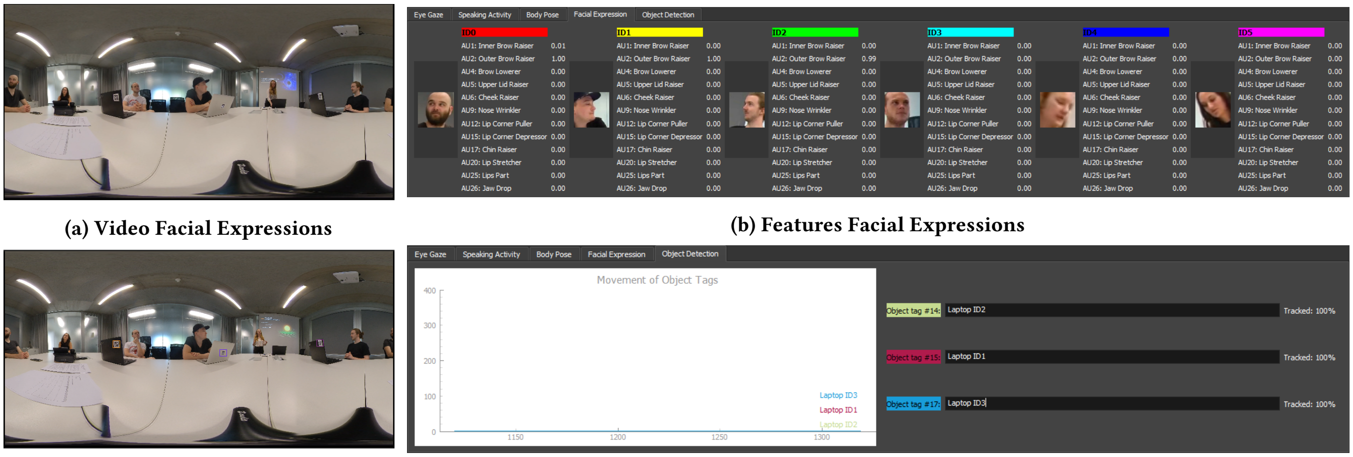
As *ConAn* can be employed in a large variety of environments, the speaker activity detection should be able to handle this variety. We therefore chose to employ the most recent, publicly available, state-of-the-art method [3] on the AVA-ActiveSpeaker dataset [79], as this dataset features a large variety of environments and speaker appearances. The method of Alcázar et al. [3] consists of two steps. In the first step, short audio snippets and corresponding individual face tracks of each potential speaker are analysed separately using CNNs. A second step models the temporal context and the relation between potential speakers using a LSTM network. We use code and pre-trained models provided by Alcázar et al. [3][6]. The face tracks are obtained from RT-GENE detections [34]. We observed that while speakers are usually assigned a higher probability than non-speakers in the softmax output of the method from Alcázar et al. [3], these probabilities are usually below $\frac{1}{2}$, leading to misclassification. To circumvent this issue, we assign an active speaker label to the user with the highest output probability. The sum of active speaker frames for each person is allocated as total speaking time and the resulting overall speaking distribution between speakers is visualized in a balance graph (see Figure 3d). Additionally, current active speakers are highlighted with a black frame.

4.6 Object Tracking

Object tracking is a complex task for which most approaches follow the tracking-by-detection scheme, where they first need to detect objects and then find the corresponding tracklets over time. As the users of *ConAn* are most likely able to define their own study procedures we decided to simplify this task by employing object tracking for pre-selected tags, as used by [58]. In particular, for our showcases we used the visual fiducial system AprilTag 2 [97],

[5] https://github.com/ZhiwenShao/PyTorch-JAANet

[6] https://github.com/fuankarion/active-speakers-context



(a) Video Facial Expressions

(b) Features Facial Expressions

	ID0	ID1	ID2	ID3	ID4	ID5
AU1: Inner Brow Raiser	0.01	0.00	0.00	0.00	0.00	0.00
AU2: Outer Brow Raiser	1.00	1.00	0.99	0.00	0.00	0.00
AU4: Brow Lowerer	0.00	0.00	0.00	0.00	0.00	0.00
AU5: Upper Lid Raiser	0.00	0.00	0.00	0.00	0.00	0.00
AU6: Cheek Raiser	0.00	0.00	0.00	0.00	0.00	0.00
AU9: Nose Wrinkler	0.00	0.00	0.00	0.00	0.00	0.00
AU12: Lip Corner Puller	0.00	0.00	0.00	0.00	0.00	0.00
AU15: Lip Corner Depressor	0.00	0.00	0.00	0.00	0.00	0.00
AU17: Chin Raiser	0.00	0.00	0.00	0.00	0.00	0.00
AU20: Lip Stretcher	0.00	0.00	0.00	0.00	0.00	0.00
AU25: Lips Part	0.00	0.00	0.00	0.00	0.00	0.00
AU26: Jaw Drop	0.00	0.00	0.00	0.00	0.00	0.00

(c) Video Object Detection

(d) Features Object Detection

Object tag #14: Laptop ID2 Tracked: 100%

Object tag #15: Laptop ID1 Tracked: 100%

Object tag #17: Laptop ID3 Tracked: 100%

Figure 4: The in depth view of the facial expression and object detection features in the view (IV) of *ConAn*.

where the tag positions are extracted with their tailored detector. For data collection, tags with a unique id are placed on the objects of interest. The detections for each frame can then be easily combined to object tracklets based on their specific tag ids. We visualize these tracklets in a dynamic object movement graph (see Figure 4d) and overlay the video with each tag position (see Figure 4c).

4.7 Export and Comparison

For further usability, we support users with a feature to export the selected sequences to a comprehensive report. The exported csv-files are split into gaze-, pose-, facial-, speaker-, and object-detection data. Additionally, all features that were displayed while using our tool, are recalculated based on the selected segments and exported as a json file. With these files users are able to compare conversation analyses of different videos side-by-side, facilitating studies including control groups, as well as longitudinal studies which require accumulating analyses over time.

5 EVALUATION

In the following, we present three use cases that highlight the variety of possible analyses, as well as potential users of *ConAn*, including researchers in the fields: HCI, social and behavioral science, and learning and education. All four 360° videos are recorded on an Insta360 One X. The videos of the three showcases are available for research purposes[7]. This will allow others to a) test the tool in depth but also to b) test new algorithms in the extraction pipeline.

5.1 General Conversation Analysis

In this video, a conversation between four participants was recorded. In order to simulate the distracting conditions of an in-the-wild exchange, the conversation took place outside and was intentionally interrupted by repeated phone calls to a single participant. After declining the first four calls, the participant answered the fifth call, upon which the recording comes to an end. This showcase highlights the disruptiveness of mobile phones in natural interactions, which has been studied mainly via interviews [89] or manual annotations [72], and is an important consideration for mobile interaction design processes [58].

After loading the data, the eye gaze tab is enabled, showing the user which participants are looking at each other in the current frame (see Figure 3b). Combining this view with *ConAn*’s other modalities, e.g. speaking activity or body movement, allows for holistic analysis of different variables related to conversation and group dynamics: in this case, the disruptiveness of mobile phones can be seen most prominently in the fact that the person receiving the phone calls had the least speaking activity. The person who looked at all the other participants the most, i.e. for 67% of the time, was also the one exhibiting the most positive facial affect (indicated by AU12 [66]) and was looked at by others only 34%, whereas the person being looked at the most (43%) only looked at other people for 18% of the time. The participant that talked the most was being looked at the least. These observations can be further compared with visual study of the interaction. For example, the person that was looked at the most had the most apparent body movement, as his hands remained on the table during the entire conversation, while the other subjects kept their hands below the table for most of the duration, and therefore exhibited less gesturing that would draw the attention of other people. However, one subject’s eye gaze was only tracked for 65% of frames, as can be directly seen within the tab (see Figure 3b), which is another important factor for consideration in the overall evaluation.

5.2 Assessment of Collaboration Quality

For assessment of group collaboration quality, we recorded a video in which three participants are engaged in a group collaboration task. In detail, the desert survival simulation [29], also known as the desert survival task, asks the participants to simulate a scenario in which they just crash-landed with a plane in the desert and need to rate 15 available items based on their importance for survival. Survival tasks are commonly used in the social and behavioral

[7] https://www.perceptualui.org/publications/penzkofer21_icmi/

sciences to analyze group behavior, such as detecting emergent leaders [82], or individual personality traits based on behavior in groups [55].

Upon loading the data in *ConAn*, the user observes in the first tab, eye gaze, that two of the three subjects have a similar amount of being watched (29% and 38%), as well as of looking at other people (36% and 25%), far exceeding the relative times for the third participant (0.12% and 12%, respectively). In the speaking activity tab, the user also observes that the two subjects additionally share the highest amount of speaking time (41% and 40%), an unsurprising finding, as in general the person speaking is most often watched by the others [66]. In the subsequent tabs, the user sees that body movement, specifically hand gesturing, and facial action units were similar for all participants. Based on these observations, the user may wish to further investigate the two subjects with the highest eye contact and speaking activity to determine emergent leadership and/or dominance as a personality trait.

5.3 Impact of Technology on Audience Behavior

The third use case comprises two videos of different presentations given by the same presenter to the same five audience members. In one presentation the audience had their powered-on laptops on the table, while in the other, everybody listened without accessing their personal devices: comparing extracted non-verbal behavior features between these videos therefore mimics a study setup with technology as the independent variable, a topic currently investigated by researchers in HCI in many different social scenarios [16, 24, 81]. At the same time, analyzing the non-verbal interaction between a presenter and his audience is frequently investigated in the field of learning and education [46, 75, 76], and *ConAn* facilitates this type of analysis by providing an export functionality, enabling the user to compare all aggregated features side by side.

In this case, the user can identify, for example, that while the ratio between being watched and looking at other people was almost equal (0.9) for the presenter in the video without devices, in the video with devices the ratio increased (1.4). In other words, in the scenario with listener devices the presenter made relatively fewer attempts to initiate eye contact with participants while being watched more in return.

6 DISCUSSION

As of today, we implemented all modality extractions and feature extractions using recent high-performing machine learning models. However, as all four domains – eye gaze, speaking activity, body pose, and facial action unit detection – are highly active areas of research, we are aware that the tool will have to be updated in the future. As this was clear from the beginning, *ConAn* has a modular extraction pipeline that allows replacing models and even add new features if they become valuable for conversation analysis in the future. This includes the possibility of making *ConAn* run in real-time when suitable algorithms become available.

The modular structure of *ConAn* also addressed the currently biggest challenge – the accuracy of the underlying extraction methods. While we use state-of-the-art machine learning models, they are not perfect yet. However, as we provide the first high-level tool on top of the latest algorithms, *ConAn* can already now drastically reduce the time investment into conversation analysis, especially for HCI researchers. Further, as we provide an easy to use GUI, *ConAn* also allows novices to conversation analysis to incorporate findings from social and behavioral science into their analyses.

Through our use cases, we highlighted the capabilities of *ConAn*. *ConAn* can help researchers in various areas, from psychological studies investigating non-verbal behavior in relation to autism spectrum disorder [80], social anxiety [84], or social phobia [9], over supporting human resources in analyzing group mood at work [12], or collective intelligence and group satisfaction [26], to investigating technology in social interactions [20, 72, 81, 89] or incorporating conversation analysis in fast, iterative design processes [58], as well as investigating the impact of non-verbal behavior on education and learning [1, 74, 75].

In contrast to other available conversation analysis tools, *ConAn* provides out-of-the-box visualizations on the group level, geared to enable insights into group dynamics at a glance. Notably, the scope of these tools varies too significantly for a quantitative comparison to yield meaningful results. Furthermore, *ConAn* is the first system that is built around the use of a 360° camera. This feature addresses one of the major challenges in group conversation analysis, i.e. the need for elaborate multi-sensor setups and the corresponding time-consuming synchronization and calibration task. Consequently, because of the portability and ease-of-use of a single 360° camera, with our tool, group conversations in any type of setting are now available for comprehensive non-verbal behavior analysis.

7 CONCLUSION AND FUTURE WORK

We presented *ConAn* – an open-source tool to perform conversation analysis using state-of-the-art machine learning models for feature extraction. *ConAn* is an easy to use tool that reduces the need for time-consuming video and audio annotation. Thus, this allows HCI researchers to quickly perform conversation analysis, for instance, during rapid prototyping to incorporate technology's impact already during design time. *ConAn* allows others to record a conversation using a single camera but retrieving a large number of features. For this, we use video and audio to extract the low-level modalities: eye gaze, speaking activity, body and hand pose, facial expressions, and information about the environment. The information-rich modalities can then be used to abstract high-level insights or even to compare multiple conversations.

In future work, we plan to conduct a user study to corroborate our findings by quantitative analysis of coder's time savings and overall user experience. In particular, we plan to compare the workflow and accuracy between our tool and manual annotations. Furthermore, it will be interesting to explore the usability of *ConAn* in extended use cases, such as video conferencing or videos with blurring for subject's privacy.

ACKNOWLEDGMENTS

P. Müller was funded by the German Ministry for Education and Research (BMBF) under grant number 01IS20075. A. Bulling was funded by the European Research Council (ERC; grant agreement 801708).

REFERENCES

[1] Karan Ahuja, Yuvraj Agarwal, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, and Amy Ogan. 2019. EduSense: Practical Classroom Sensing at Scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (Sept. 2019), 1–26. https://doi.org/10.1145/3351229

[2] Karan Ahuja, Chris Harrison, Mayank Goel, and Robert Xiao. 2019. MeCap: Whole-Body Digitization for Low-Cost VR/AR Headsets. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 453–462. https://doi.org/10.1145/3332165.3347889

[3] Juan León Alcázar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbeláz, and Bernard Ghanem. 2020. Active Speakers in Context. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '20)*. IEEE, 12462–12471. https://doi.org/10.1109/CVPR42600.2020.01248

[4] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *OpenFace: A general-purpose face recognition library with mobile applications.* Technical Report. CMU-CS-16-118, CMU School of Computer Science.

[5] Keith Anderson, Elisabeth André, Tobias Baur, Sara Bernardini, Mathieu Chollet, Evi Chryssafidou, Ionut Damian, Cathy Ennis, Arjan Egges, Patrick Gebhard, et al. 2013. The TARDIS Framework: Intelligent Virtual Agents for Social Coaching in Job Interviews. In *International Conference on Advances in Computer Entertainment Technology*. Springer International Publishing, Cham, 476–491. https://doi.org/10.1007/978-3-319-03161-3_35

[6] Marc H. Anderson and Peter Y. T. Sun. 2017. Reviewing Leadership Styles: Overlaps and the Need for a New 'Full-Range' Theory. *International Journal of Management Reviews* 19, 1 (2017), 76–96. https://doi.org/10.1111/ijmr.12082

[7] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 2 (2012), 356–370. https://doi.org/10.1109/TASL.2011.2125954

[8] Michael Argyle and Janet Dean. 1965. Eye-Contact, Distance and Affiliation. *Sociometry* 28, 3 (1965), 289–304. https://doi.org/10.2307/2786027

[9] Sarah R. Baker and Robert J. Edelmann. 2002. Is social phobia related to lack of social skills? Duration of skill-related behaviours and ratings of behavioural adequacy. *British Journal of Clinical Psychology* 41, 3 (2002), 243–257. https://doi.org/10.1348/014466502760379118

[10] Yutong Ban, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. 2019. Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers. *arXiv:1809.10961 [cs, stat]* (Oct. 2019). arXiv:1809.10961 [cs] http://arxiv.org/abs/1809.10961

[11] Sigal G. Barsade. 2002. The Ripple Effect: Emotional Contagion and its Influence on Group Behavior:. *Administrative Science Quarterly* (2002). https://doi.org/10.2307/3094912

[12] Caroline A. Bartel and Richard Saavedra. 2000. The Collective Construction of Work Group Moods. *Administrative Science Quarterly* 45, 2 (June 2000), 197. https://doi.org/10.2307/2667070

[13] Nirit Bauminger. 2002. The Facilitation of Social-Emotional Understanding and Social Interaction in High-Functioning Children with Autism: Intervention Outcomes. *Journal of Autism and Developmental Disorders* 32, 4 (Aug. 2002), 283–298. https://doi.org/10.1023/A:1016378718278

[14] Tobias Baur, Ionut Damian, Florian Lingenfelser, Johannes Wagner, and Elisabeth André. 2013. Nova: Automated analysis of nonverbal signals in social interactions. In *International Workshop on Human Behavior Understanding*. Springer, 160–171. https://doi.org/10.1007/978-3-319-02714-2_14

[15] Janet Beavin Bavelas and Nicole Chovil. 2000. Visible Acts of Meaning: An Integrated Message Model of Language in Face-to-Face Dialogue. *Journal of Language and Social Psychology* 19, 2 (2000), 163–194. https://doi.org/10.1177/0261927X00019002001

[16] 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 *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction* (Santa Monica, California) *(Gaze-In '12)*. Association for Computing Machinery, New York, NY, USA, Article 10, 6 pages. https://doi.org/10.1145/2401836.2401846

[17] Abdallah Benzine, Florian Chabot, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. 2020. PandaNet: Anchor-Based Single-Shot Multi-Person 3D Pose Estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR '20)*. IEEE Computer Society, Los Alamitos, CA, USA, 6855–6864. https://doi.org/10.1109/CVPR42600.2020.00689

[18] Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. 2017. Prediction of the leadership style of an emergent leader using audio and visual nonverbal features. *IEEE Transactions on Multimedia* 20, 2 (2017), 441–456. https://doi.org/10.1109/TMM.2017.2740062

[19] Cigdem Beyan, Nicolò Carissimi, Francesca Capozzi, Sebastiano Vascon, Matteo Bustreo, Antonio Pierro, Cristina Becchio, and Vittorio Murino. 2016. Detecting emergent leader in a meeting environment using nonverbal visual features only. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16)*. Association for Computing Machinery, Tokyo, Japan, 317–324. https://doi.org/10.1145/2993148.2993175

[20] Barry Brown, Moira McGregor, and Donald McMillan. 2015. Searchable Objects: Search in Everyday Conversation. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 508–517. https://doi.org/10.1145/2675133.2675206

[21] George Butterworth and Shoji Itakura. 2000. How the eyes, head and hand serve definite reference. *British Journal of Developmental Psychology* 18, 1 (2000), 25–50. https://doi.org/10.1348/026151000165553

[22] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), 1–1. https://doi.org/10.1109/TPAMI.2019.2929257

[23] Joseph N. Cappella and Sally Planalp. 1981. Talk and Silence Sequences in Informal Conversations III: Interspeaker Influence. *Human Communication Research* 7, 2 (Dec. 1981), 117–132. https://doi.org/10.1111/j.1468-2958.1981.tb00564.x

[24] Debaleena Chattopadhyay, Kenton O'Hara, Sean Rintel, and Roman Rädle. 2016. Office Social: Presentation Interactivity for Nearby Devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2487–2491. https://doi.org/10.1145/2858036.2858337

[25] Lei Chen, Gary Feng, Jilliam Joe, Chee Wee Leong, Christopher Kitchen, and Chong Min Lee. 2014. Towards Automated Assessment of Public Speaking Skills Using Multimodal Cues. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14)*. Association for Computing Machinery, New York, NY, USA, 200–203. https://doi.org/10.1145/2663204.2663265

[26] Prerna Chikersal, Maria Tomprou, Young Ji Kim, Anita Williams Woolley, and Laura Dabbish. 2017. Deep Structures of Collaboration: Physiological Correlates of Collective Intelligence and Group Satisfaction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 873–888. https://doi.org/10.1145/2998181.2998250

[27] Ross Cutler, Ramin Mehran, Sam Johnson, Cha Zhang, Adam Kirk, Oliver Whyte, and Adarsh Kowdle. 2020. Multimodal active speaker detection and virtual cinematography for video conferencing. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '20)*. IEEE, 4527–4531. https://doi.org/10.1109/ICASSP40776.2020.9053171

[28] Dipankar Das, Yoshinori Kobayashi, and Yoshinori Kuno. 2013. Attracting attention and establishing a communication channel based on the level of visual focus of attention. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2194–2201. https://doi.org/10.1109/IROS.2013.6696663

[29] Brad Deacon. 2016. EFL Students in the Desert: Using Survival Simulations to Improve Teamwork.

[30] Ap Dijksterhuis and John A. Bargh. 2001. The perception-behavior expressway: Automatic effects of social perception on social behavior. In *Advances in Experimental Social Psychology*. Vol. 33. Elsevier, 1–40. https://doi.org/10.1016/S0065-2601(01)80003-4

[31] Paul Ekman and Wallace V. Friesen. 1976. Measuring facial movement. *Environmental psychology and nonverbal behavior* 1, 1 (1976), 56–75. https://doi.org/10.1007/BF01115465

[32] David Engel, Anita Williams Woolley, Ishani Aggarwal, Christopher F. Chabris, Masamichi Takahashi, Keiichi Nemoto, Carolin Kaiser, Young Ji Kim, and Thomas W. Malone. 2015. Collective Intelligence in Computer-Mediated Collaboration Emerges in Different Contexts and Cultures. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, Seoul, Republic of Korea, 3769–3778. https://doi.org/10.1145/2702123.2702259

[33] Pierre Feyereisen and Jacques-Dominique De Lannoy. 1991. *Gestures and speech: Psychological investigations.* Cambridge University Press.

[34] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. 2018. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In *The European Conference on Computer Vision (ECCV '18)*. Springer International Publishing, 339–357. https://doi.org/10.1007/978-3-030-01249-6_21

[35] Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree. 2017. Speaker diarization using deep neural network embeddings. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '17)*. IEEE, 4930–4934. https://doi.org/10.1109/ICASSP.2017.7953094

[36] Binyam Gebrekidan Gebre, Peter Wittenburg, and Tom Heskes. 2013. The gesturer is the speaker. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3751–3755. https://doi.org/10.1109/ICASSP.2013.6638359

[37] Israel D. Gebru, Silèye Ba, Xiaofei Li, and Radu Horaud. 2018. Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 5 (May 2018), 1086–1099. https://doi.org/10.1109/TPAMI.2017.2648793 arXiv:1603.09725

[38] John B Haviland. 2003. How to point in Zinacantán. *Pointing: Where language, culture, and cognition meet* 139169 (2003).

[39] Alexander Heimerl, Tobias Baur, Florian Lingenfelser, Johannes Wagner, and Elisabeth André. 2019. NOVA-a tool for eXplainable Cooperative Machine Learning. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 109–115.

[40] Arjan van Hessen, Silvia Calamai, Henk van den Heuvel, Stefania Scagliola, Norah Karrouche, Jeannine Beeken, Louise Corti, and Christoph Draxler. 2020. Speech, Voice, Text, and Meaning: A Multidisciplinary Approach to Interview Data through the use of digital tools. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 886–887. https://doi.org/10.1145/3382507.3420054

[41] Simon Ho, Tom Foulsham, and Alan Kingstone. 2015. Speaking and listening with the eyes: gaze signaling during dyadic interactions. *PloS one* 10, 8 (2015), e0136905. https://doi.org/10.1371/journal.pone.0136905

[42] Sabrina Hoppe, Tobias Loetscher, Stephanie Morey, and Andreas Bulling. 2018. Eye movements during everyday behavior predict personality traits. *Frontiers in Human Neuroscience* 12 (2018), 105:1–105:8. https://doi.org/10.3389/fnhum.2018.00105

[43] Hiroshi Ishii and Minoru Kobayashi. 1992. ClearBoard: A Seamless Medium for Shared Drawing and Conversation with Eye Contact. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, USA) *(CHI '92)*. Association for Computing Machinery, New York, NY, USA, 525–532. https://doi.org/10.1145/142750.142977

[44] J. Jesna, Athi S. Narayanan, and Kamal Bijlani. 2018. Automatic Hand Raise Detection by Analyzing the Edge Structures. In *Emerging Research in Computing, Information, Communication and Applications*. Springer Singapore, Singapore, 171–180. https://doi.org/10.1007/978-981-10-4741-1_16

[45] Kristiina Jokinen, Masafumi Nishida, and Seiichi Yamamoto. 2010. On eye-gaze and turn-taking. In *Proceedings of the 2010 workshop on Eye gaze in intelligent human machine interaction (EGIHMI '10)*. Association for Computing Machinery, New York, NY, USA, 118–123. https://doi.org/10.1145/2002333.2002352

[46] L. Thomas Keith, Louis G. Tornatzky, and L. Eudora Pettigrew. 1974. An analysis of verbal and nonverbal classroom teaching behaviors. *The Journal of Experimental Education* 42, 4 (1974), 30–38. https://doi.org/10.1080/00220973.1974.11011490

[47] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. 2019. Gaze360: Physically Unconstrained Gaze Estimation in the Wild. In *The IEEE International Conference on Computer Vision (ICCV '19)*. IEEE. https://doi.org/10.1109/ICCV.2019.00701

[48] Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica* 26 (1967), 22–63. https://doi.org/10.1016/0001-6918(67)90005-4

[49] Taemie Kim, Agnes Chang, Lindsey Holland, and Alex Sandy Pentland. 2008. Meeting mediator: enhancing group collaborationusing sociometric feedback. In *Proceedings of the ACM 2008 conference on Computer supported cooperative work - CSCW '08*. ACM Press, San Diego, CA, USA, 457. https://doi.org/10.1145/1460563.1460636

[50] Kasia Kozlowska, Kerri J. Brown, Donna M. Palmer, and Lea M. Williams. 2013. Specific Biases for Identifying Facial Expression of Emotion in Children and Adolescents With Conversion Disorders. *Psychosomatic Medicine* 75, 3 (2013). https://doi.org/10.1097/PSY.0b013e318286be43

[51] Bruno Lepri, Ramanathan Subramanian, Kyriaki Kalimeri, Jacopo Staiano, Fabio Pianesi, and Nicu Sebe. 2010. Employing social gaze and speaking activity for automatic determination of the *Extraversion* trait. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI '10)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/1891903.1891913

[52] Jing Li, Yang Mi, Gongfa Li, and Zhaojie Ju. 2019. CNN-Based Facial Expression Recognition from Annotated RGB-D Images for Human–Robot Interaction. *International Journal of Humanoid Robotics* 16, 04 (2019), 1941002. https://doi.org/10.1142/S0219843619410020

[53] Filip Lievens and Richard J. Klimoski. 2001. Understanding the assessment center process: Where are we now? *International Review of Industrial and Organizational Psychology* 16 (April 2001), 245–286. https://ink.library.smu.edu.sg/lkcsb_research/5823

[54] Chen Ma, Li Chen, and Junhai Yong. 2019. AU R-CNN: Encoding expert prior knowledge into R-CNN for action unit detection. *Neurocomputing* 355 (2019), 35 – 47. https://doi.org/10.1016/j.neucom.2019.03.082

[55] Nadia Mana, Bruno Lepri, Paul Chippendale, Alessandro Cappelletti, Fabio Pianesi, Piergiorgio Svaizer, and Massimo Zancanaro. 2007. Multimodal Corpus of Multi-Party Meetings for Automatic Social Behavior Analysis and Personality Traits Detection. In *Proceedings of the 2007 Workshop on Tagging, Mining and Retrieval of Human Related Activity Information* (Nagoya, Japan) *(TMR '07)*. Association for Computing Machinery, New York, NY, USA, 9–14. https://doi.org/10.1145/1330588.1330590

[56] Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, and Philip Trinh. 2012. Automatic detection of non-posed facial action units. In *2012 19th IEEE International Conference on Image Processing*. IEEE, 1817–1820. https://doi.org/10.1109/ICIP.2012.6467235

[57] S. Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. 2013. DISFA: A Spontaneous Facial Action Intensity Database. *IEEE Transactions on Affective Computing* 4, 2 (2013), 151–160. https://doi.org/10.1109/T-AFFC.2013.4

[58] Sven Mayer, Lars Lischke, Paweł W. Woźniak, and Niels Henze. 2018. Evaluating the Disruptiveness of Mobile Interactions: A Mixed-Method Approach. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173980

[59] Sven Mayer, Jens Reinhardt, Robin Schweigert, Brighten Jelke, Valentin Schwind, Katrin Wolf, and Niels Henze. 2020. Improving Humans' Ability to Interpret Deictic Gestures in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376340

[60] Daniel McDuff, Kael Rowan, Piali Choudhury, Jessica Wolk, ThuVan Pham, and Mary Czerwinski. 2019. A multimodal emotion sensing platform for building emotion-aware applications. *arXiv preprint arXiv:1903.12133* (2019). arXiv:1903.12133

[61] AJung Moon, Daniel M. Troniak, Brian Gleeson, Matthew K.X.J. Pan, Minhua Zheng, Benjamin A. Blumer, Karon MacLean, and Elizabeth A. Croft. 2014. Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction (HRI '14)*. Association for Computing Machinery, New York, NY, USA, 334–341. https://doi.org/10.1145/2559636.2559656

[62] Philipp Müller, Sikandar Amin, Prateek Verma, Mykhaylo Andriluka, and Andreas Bulling. 2015. Emotion recognition from embedded bodily expressions and speech during dyadic interactions. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII '15)*. IEEE, 663–669. https://doi.org/10.1109/ACII.2015.7344640

[63] Philipp Müller and Andreas Bulling. 2019. Emergent Leadership Detection Across Datasets. In *2019 International Conference on Multimodal Interaction* (Suzhou, China) *(ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 274–278. https://doi.org/10.1145/3340555.3353721 arXiv:1905.02058 [cs]

[64] 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 *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (Warsaw, Poland) *(ETRA '18)*. Association for Computing Machinery, New York, NY, USA, Article 31, 10 pages. https://doi.org/10.1145/3204493.3204549

[65] Philipp Müller, Ekta Sood, and Andreas Bulling. 2020. Anticipating Averted Gaze in Dyadic Interactions. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20)*. Association for Computing Machinery, New York, NY, USA, Article 7, 10 pages. https://doi.org/10.1145/3379155.3391332

[66] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 153–164. https://doi.org/10.1145/3172944.3172969

[67] Magalie Ochs, Nathan Libermann, Axel Boidin, and Thierry Chaminade. 2017. Do you speak to a human or a virtual agent? automatic analysis of user & social cues during mediated communication. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI '17)*. Association for Computing Machinery, New York, NY, USA, 197–205. https://doi.org/10.1145/3136755.3136807

[68] Eyal Ofek, Shamsi T. Iqbal, and Karin Strauss. 2013. Reducing Disruption from Subtle Information Delivery during a Conversation: Mode and Bandwidth Investigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 3111–3120. https://doi.org/10.1145/2470654.2466425

[69] 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* (Seattle, Washington, USA) *(ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 15–22. https://doi.org/10.1145/2818346.2820757

[70] Dina G. Okamoto, Lisa Slattery Rashotte, and Lynn Smith-Lovin. 2002. Measuring Interruption: Syntactic and Contextual Methods of Coding Conversation. *Social Psychology Quarterly* 65, 1 (2002), 38–55. https://doi.org/10.2307/3090167

[71] Ulrich J. Pfeiffer, Kai Vogeley, and Leonhard Schilbach. 2013. From gaze cueing to dual eye-tracking: Novel approaches to investigate the neural correlates of gaze in social interaction. *Neuroscience & Biobehavioral Reviews* 37, 10, Part 2 (2013), 2516 – 2528. https://doi.org/10.1016/j.neubiorev.2013.07.017

[72] Martin Porcheron, Joel E. Fischer, and Sarah Sharples. 2016. Using Mobile Phones in Pub Talk. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1649–1661.

https://doi.org/10.1145/2818048.2820014

[73] Luis P. Prieto, Kshitij Sharma, Pierre Dillenbourg, and María Jesús. 2016. Teaching analytics: towards automatic extraction of orchestration graphs using wearable sensors. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16)*. Association for Computing Machinery, Edinburgh, United Kingdom, 148–157. https://doi.org/10.1145/2883851.2883927

[74] Mirko Raca and Pierre Dillenbourg. 2013. System for assessing classroom attention. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK '13)*. Association for Computing Machinery, Leuven, Belgium, 265–269. https://doi.org/10.1145/2460296.2460351

[75] Mirko Raca, Lukasz Kidzinski, and Pierre Dillenbourg. 2015. Translating head motion into attention-towards processing of student's body-language. In *Proceedings of the 8th International Conference on Educational Data Mining*. Madrid, Spain.

[76] Kelly A. Rocca. 2010. Student Participation in the College Classroom: An Extended Multidisciplinary Literature Review. *Communication Education* 59, 2 (April 2010), 185–213. https://doi.org/10.1080/03634520903505936

[77] Natalia Stewart Rosenfield, Kathleen Lamkin, Jennifer Re, Kendra Day, LouAnne Boyd, and Erik Linstead. 2019. A Virtual Reality System for Practicing Conversation Skills for Children with Autism. *Multimodal Technologies and Interaction* 3, 2 (Apr 2019), 28. https://doi.org/10.3390/mti3020028

[78] Federico Rossano. 2013. Gaze in Conversation. *The handbook of conversation analysis* (2013), 308.

[79] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, and Caroline Pantofaru. 2020. Ava Active Speaker: An Audio-Visual Dataset for Active Speaker Detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '20)*. IEEE, 4492–4496. https://doi.org/10.1109/ICASSP40776.2020.9053900

[80] Agata Rozga, Ted Hutman, Gregory S Young, Sally J Rogers, Sally Ozonoff, Mirella Dapretto, and Marian Sigman. 2011. Behavioral profiles of affected and unaffected siblings of children with autism: Contribution of measures of mother–infant interaction and nonverbal communication. *Journal of autism and developmental disorders* 41, 3 (2011), 287–301. https://doi.org/10.1007/s10803-010-1051-6

[81] Tarja Salmela, Ashley Colley, and Jonna Häkkilä. 2019. Together in Bed? Couples' Mobile Technology Use in Bed. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300732

[82] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. 2012. A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups. *IEEE Transactions on Multimedia* 14, 3 (June 2012), 816–832. https://doi.org/10.1109/TMM.2011.2181941

[83] Alexander Schafer, Tomoko Isomura, Gerd Reis, Katsumi Watanabe, and Didier Stricker. 2020. MutualEyeContact: A Conversation Analysis Tool with Focus on Eye Contact. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20)*. Association for Computing Machinery, New York, NY, USA, Article 1, 5 pages. https://doi.org/10.1145/3379156.3391340

[84] Franklin R. Schneier, Thomas L. Rodebaugh, Carlos Blanco, Hillary Lewin, and Michael R. Liebowitz. 2011. Fear and avoidance of eye contact in social anxiety disorder. *Comprehensive Psychiatry* 52, 1 (Jan. 2011), 81–87. https://doi.org/10.1016/j.comppsych.2010.04.006

[85] Muhammad Shahid, Cigdem Beyan, and Vittorio Murino. 2019. Voice Activity Detection by Upper Body Motion Analysis and Unsupervised Domain Adaptation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW '19)*. IEEE, 0–0. https://doi.org/10.1109/ICCVW.2019.00159

[86] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. 2020. JAA-Net: Joint Facial Action Unit Detection and Face Alignment Via Adaptive Attention. *International Journal of Computer Vision* (2020). https://doi.org/10.1007/s11263-020-01378-z

[87] Kalin Stefanov, Baiyu Huang, Zongjian Li, and Mohammad Soleymani. 2020. OpenSense: A Platform for Multimodal Data Acquisition and Behavior Perception. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 660–664. https://doi.org/10.1145/3382507.3418832

[88] Giota Stratou and Louis-Philippe Morency. 2017. MultiSense—Context-aware nonverbal behavior analysis framework: A psychological distress use case. *IEEE Transactions on Affective Computing* 8, 2 (2017), 190–203. https://doi.org/10.1109/TAFFC.2016.2614300

[89] Norman Makoto Su and Lulu Wang. 2015. From Third to Surveilled Place: The Mobile in Irish Pubs. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1659–1668. https://doi.org/10.1145/2702123.2702574

[90] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR '19)*. IEEE, 5693–5703. https://doi.org/10.1109/CVPR.2019.00584

[91] Ashish Kumar Tiwari. 2015. Non-verbal communication-an essence of interpersonal relationship at workplace. *Management Insight* 11, 2 (2015), 109–114. http://journals.smsvaranasi.com/index.php/managementinsight/article/view/431

[92] Alexander Toshev and Christian Szegedy. 2014. DeepPose: Human Pose Estimation via Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*. IEEE. https://doi.org/10.1109/CVPR.2014.214

[93] Laura Valenzeno, Martha W. Alibali, and Roberta Klatzky. 2003. Teachers' gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology* 28, 2 (April 2003), 187–204. https://doi.org/10.1016/S0361-476X(02)00007-3

[94] Alex B. Van Zant and Jonah Berger. 2020. How the voice persuades. *Journal of Personality and Social Psychology* 118, 4 (April 2020), 661–682. https://doi.org/10.1037/pspi0000193

[95] Ulrich von Zadow and Raimund Dachselt. 2017. GIAnT: Visualizing Group Interaction at Large Wall Displays. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 2639–2647. https://doi.org/10.1145/3025453.3026006 event-place: Denver, Colorado, USA.

[96] Johannes Wagner, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth André. 2013. The Social Signal Interpretation (SSI) Framework: Multimodal Signal Processing and Recognition in Real-Time. In *Proceedings of the 21st ACM International Conference on Multimedia* (Barcelona, Spain) *(MM '13)*. Association for Computing Machinery, New York, NY, USA, 831–834. https://doi.org/10.1145/2502081.2502223

[97] John Wang and Edwin Olson. 2016. AprilTag 2: Efficient and robust fiducial detection. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4193–4198. https://doi.org/10.1109/IROS.2016.7759617

[98] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. 2018. Speaker Diarization with LSTM. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5239–5243. https://doi.org/10.1109/ICASSP.2018.8462628

[99] Wei Wei, Qingxuan Jia, and Gang Chen. 2016. Real-time facial expression recognition for affective computing based on Kinect. In *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA '16)*. IEEE, 161–165. https://doi.org/10.1109/ICIEA.2016.7603570

[100] Zachary Witkower, Jessica L. Tracy, Joey T. Cheng, and Joseph Henrich. 2020. Two signals of social rank: Prestige and dominance are associated with distinct nonverbal displays. *Journal of Personality and Social Psychology* 118, 1 (2020), 89–120. https://doi.org/10.1037/pspi0000181

[101] Jacob O. Wobbrock. 2019. *Situationally-Induced Impairments and Disabilities*. Springer London, London, 59–92. https://doi.org/10.1007/978-1-4471-7440-0_5

[102] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* 330, 6004 (Oct. 2010), 686–688. https://doi.org/10.1126/science.1193147

[103] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2019. Evaluation of Appearance-Based Methods and Implications for Gaze-Based Applications. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300646