ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild

Chirag Raman[1*] **Jose Vargas-Quiros**[1*] **Stephanie Tan**[1*] **Ashraful Islam**[2]

Ekin Gedik[1] **Hayley Hung**[1]

[1]Delft University of Technology, Delft, The Netherlands
{c.a.raman, j.d.vargasquiros, s.tan-1, e.gedik, h.hung}@tudelft.nl

[2]Rensselaer Polytechnic Institute, New York, USA
islama6@rpi.edu

Abstract

Recording the dynamics of unscripted human interactions in the wild is challenging due to the delicate trade-offs between several factors: participant privacy, ecological validity, data fidelity, and logistical overheads. To address these, following a *datasets for the community by the community* ethos, we propose the Conference Living Lab (ConfLab): a new concept for multimodal multisensor data collection of in-the-wild free-standing social conversations. For the first instantiation of ConfLab described here, we organized a real-life professional networking event at a major international conference. Involving 48 conference attendees, the dataset captures a diverse mix of status, acquaintance, and networking motivations. Our capture setup improves upon the data fidelity of prior in-the-wild datasets while retaining privacy sensitivity: 8 videos (1920×1080, 60 fps) from a non-invasive overhead view, and custom wearable sensors with onboard recording of body motion (full 9-axis IMU), privacy-preserving low-frequency audio (1250 Hz), and Bluetooth-based proximity. Additionally, we developed custom solutions for distributed hardware synchronization at acquisition, and time-efficient continuous annotation of body keypoints and actions at high sampling rates. Our benchmarks showcase some of the open research tasks related to in-the-wild privacy-preserving social data analysis: keypoints detection from overhead camera views, skeleton-based no-audio speaker detection, and F-formation detection.

1 Introduction

A crucial challenge towards developing artificial socially intelligent systems is understanding how *real-life* situational contexts affect social human behavior [1]. Social-science findings indeed show that the dynamics of how we conduct daily interactions vary significantly depending on the social situation [2–4]. Unfortunately, such dynamics are not adequately captured by many data collection setups where role-played or scripted scenarios are typical [5].

In this paper we address the problem of collecting a privacy-sensitive dataset of unscripted social dynamics of real-life relationships where encounters can influence someone's daily life. We argue that doing so requires recording these exchanges in the natural ecology, requiring an approach

[*]Equal contribution

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

Figure 1: Snapshot of the interaction area from our cameras. We annotated only cameras highlighted with red borders (high scene overlap). For a clearer visual impression of the scene, we omit cameras 1 (few people recorded) and 5 (failed early in the event). Faces blurred to preserve privacy.

different from the typical setup of locally-organized studies. Specifically, we focus on free-standing interactions within the setting of an international conference (see Figure 1).

Recording an international community in its natural habitat is characterized by several intersecting challenges: an intrinsic trade-off exists between data fidelity, ecological validity, and privacy preservation. For ecological validity, a non-invasive capture setup is essential for mitigating any influence on behavior naturalness [6–8].The most common solution involves mounting cameras from aerial perspectives such as top-down [9, 10] and elevated-side views [11–13]. Now elevated-side views make it easy to capture sensitive personal information such as faces, which leads to several ethical concerns. For instance, capturing faces has been related to harmful downstream surveillance applications [14]. Besides, state-of-the-art (SOTA) body-keypoint estimation techniques perform poorly on aerial perspectives [9, 15], making the extraction of automatic pose annotations challenging (Figure 3). To avoid such issues, some researchers have turned to more privacy-preserving wearable sensors shown to benefit many behavior analysis tasks [8, 16, 17].

In all, the closest related datasets (see Table 1) suffer from several technical limitations precluding the analysis and modeling of fine-grained social behavior: (i) lack of articulated pose annotations; (ii) a limited number of people in the scene, preventing complex interactions such as group splitting/merging behaviors, and (iii) an inadequate data sampling-rate and synchronization-latency to study time-sensitive social phenomena [18, Sec. 3.3]. To address all these limitations, we propose the Conference Living Lab (ConfLab): a new concept for multimodal multisensor data collection of ecologically-valid social settings. From the first instantiation of ConfLab, we provide a high-fidelity dataset of 48 participants at a professional networking event.

Methodological Contributions: We describe a data collection design that captures a diverse mix of real levels of seniority, acquaintance, affiliation, and motivation to network (see Figure 2). This was achieved by organizing ConfLab as part of a major international scientific conference. ConfLab had these goals: (i) a data collection effort follwing a *by the community for the community* ethos: the more volunteers, the more data, (ii) volunteers who potentially use the data can experience first-hand potential privacy and ethical considerations related to sharing their own data, (iii) in light of recent data sourcing issues [14, 20], we incorporated privacy and invasiveness considerations directly into the decision-making process regarding sensor type, positioning, and sample-rates.

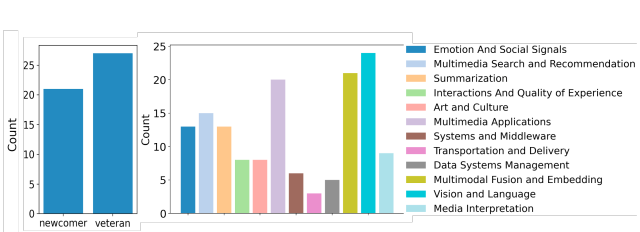


Figure 2: Frequency of newcomer/veteran participants (left) and reported research interests (right).

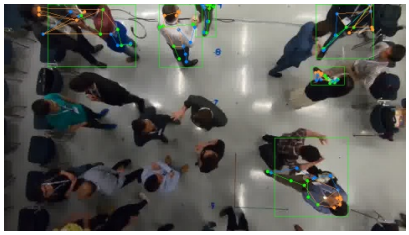


Figure 3: Keypoint detection using pretrained RSN [19]. Additional SOTA results are in Appendix F.1

Table 1: Comparison of ConfLab with prior datasets of free-standing conversation groups in in-the-wild social interaction settings. Conflab is the first and only social interaction dataset that offers skeletal keypoints and speaking status at high annotation resolution, as well as hardware synchronized camera and multimodal wearable signals at high resolution.

Dataset	People/ Scene	Video	Manual Annotations	Wearable Signals	Synchronization
Cocktail [13]†	7	512×384	F-formations (20 and 30 min, $1/5$ Hz)	None	Unknown
CoffeeBreak [12]	14	1440×1080	F-formations (130 frames in two sequences)	None	None
IDIAP [10]	> 50	180 min; 654×439 20 fps	F-formations (82 independent frames)	None	None
SALSA [11]†	18	60 min; 1024×768 15 fps	Bounding boxes (30 min) Head & body ori. (30 min) F-formations (60 min) (all $1/3$ Hz)	Audio MFCCs (30 Hz) Acceleration (20 Hz) IR proximity (1 Hz)	Post-hoc infra-red event-based (no-drift assumption)
MnM [9]†	32	30 min; 1920×1080 30 fps	Bounding boxes (30 min, 1 Hz ‡) F-formations (10 min, 1 Hz) Actions (45 min, 1 Hz‡)	Accelerometer (20 Hz) Radio proximity (1 Hz)	Intra-wearable sync via gossiping protocol; Inter-modal sync using manual inspection @1 Hz
ConfLab	48	$\sim$ 45 min; 1920×1080 **60 fps**	**17 keypoints (16 min, 60 Hz)** F-formations (16 min, 1 Hz) **Speaking status (16 min, 60 Hz)**	**Low-freq. audio (1250 Hz)** **BT proximity (5 Hz)** **9-axis IMU (56 Hz)**	**Wireless hardware sync at acquisition, max latency of $\sim$ 13 ms** [18]

† Includes self-assessed personality ratings ‡ Upsampled to 20 Hz using Vatic [25] BT: Bluetooth IMU: Inertial Measurement Unit

Technical Contributions: **(i) aerial-view articulated pose**: our annotations of 17 full-body keypoints enable improvements in (a) pose estimation and tracking, (b) pose-based recognition of social actions (under-explored in the top-down perspective), (c) pose-based F-formation estimation (has not been possible from prior work [10, 21–23]), and (d) the direct study of interaction dynamics using full body poses (previously limited to lab settings [24]). **(ii) subtle body dynamics**: we are the first to use a full 9-axis Inertial Measurement Unit (IMU) enabling a richer representation of behaviour at higher sample rates; previous rates were found to be insufficient for downstream tasks [17]. **(iii) enabling finer temporal-scale research questions**: a sub-second crossmodal latency of $\sim$ 13 ms along with higher sampling rate of features (60 fps video, 56 Hz IMU) opens the gateway for the in-the-wild study of nuanced time-sensitive social behaviors like mimicry and synchrony.

2 Related Work

Early datasets of in-the-wild social events either spanned only a few minutes (e.g. Coffee Break [12]), or were recorded at such a large distance from the participants that performing robust, automated person detection or tracking with SOTA approaches was non-trivial (e.g. Idiap Poster Data [10]). More recently, two different strategies have emerged to circumvent such issues.

One approach involves fully instrumented labs with a high resolution multi-camera setup for video and audio data. Here automatic detectors [24, 26, 27] could be applied to obtain poses. This circumvents the cost- and labor-intensive process of manually labeling head poses, at the cost of less portable sensing setups. Notable examples of such in-the-lab studies include seated scenarios, such as the AMI meeting corpus [28], and more recently standing scenarios like the Panoptic Dataset [24]. Both enable the learning of multimodal behavioral dynamics. However, the dynamics of seated, scripted, or role-playing scenarios are different from that of an unconstrained social setting such as ours. In contrast, ConfLab moves out of the lab with a more modular and portable multimodal, multisensor solution that scales easily in the wild.

Another approach exploited wearable sensor data to allow for multimodal processing—sensors included 3 or 6 DOF inertial measurement units (IMU); infrared, bluetooth, or radio sensors to measure proximity; or microphones for speech behavior [9, 11]. While proximity has been used as a proxy of face-to-face interaction [11, 29–32], recent findings highlight significant problems with such an assumption [33]. Such errors can have a significant impact on the machine-perceived experience of an individual, precluding the development of personalized technology. Chalcedony badges used by

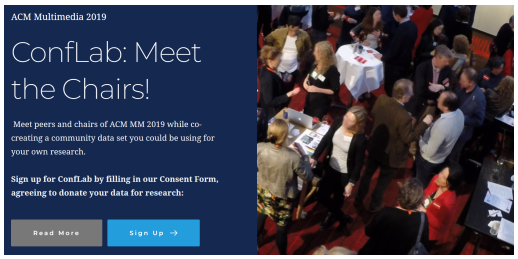


Figure 4: Screenshots from the *ConfLab: Meet the Chairs!* event website

Figure 5: The Midge

[9] show more promising results with a radio-based proximity sensor and accelerometer [34], but such data remains insufficient for more downstream tasks due to the relatively low sample (20Hz) and annotation (1Hz) frequency [17]. In light of these challenges in wearable sensing, ConfLab features custom-developed Midge sensors that enable more flexible and fine-grained on-device recording. At the same time, ConfLab enables researchers in the wearable and ubiquitous computing communities to investigate the benefit of exploiting wearable and multimodal data.

Furthermore, while both SALSA [11] and MatchNMingle [9] capture a multimodal dataset of a large group of individuals involved in mingling behavior, the inter-modal synchronization is only guaranteed at 1/3 Hz and 1 Hz, respectively. Prior works coped with lower tolerances by computing summary statistics over input windows [17, 35, 36]. While 1 Hz is able to capture some conversation dynamics [37], it is insufficient to study fine-grained social phenomena such as back-channeling or mimicry that involve far lower latencies [18, Sec. 3.3]. ConfLab provides data streams with higher sampling rates, synchronized at acquisition with our method shown to yield a 13 ms latency at worst [18] (see Sec. 3). Table 1 summarizes the differences between ConfLab and other related datasets.

3 Data Acquisition

In this section we describe the considerations, design, and supporting community engagement activities for the first instantiation of ConfLab at ACM Multimedia 2019 (MM'19), to serve as a template and case study for other similar efforts.

Ecological Validity and Recruitment An often-overlooked but crucial aspect of in-the-wild data collection is the design and ecological validity of the interaction setting [6–8]. To capture natural interactions in a professional setting and encourage mixed levels of status, acquaintance, and motivations to network, we co-designed a networking event with the MM'19 organizers called *Meet the Chairs!* Our event website (`https://conflab.ewi.tudelft.nl/`) served to inform participants about the goals of a community created dataset, and transparently describe the data collection process (Figure 4). During the conference, participants were recruited via word-of-mouth marketing, social media, conference announcements, and the event website. As an additional incentive beyond interacting with the Chairs and participating in a community-driven data endeavor, we provided attendees with post-hoc insights into their networking behavior from the collected wearable-sensors data. See Supplementary material for a sample participant report.

Privacy and Ethics The collection and sharing of ConfLab is GDPR compliant. The dataset design and process was approved by both, the Human Research Ethics Committee (HREC) at our institution (TUDelft) and the conference location's national authorities (France). All participants gave consent for the recording and sharing of their data at registration.(See the Datasheet in the Appendix for the consent form.) Given the involvement of private human data, ConfLab is only available for academic research purposes under an End User License Agreement. Such an *as open as possible and as closed as necessary* ethos for open science acknowledges the limitation that personal data places on open sharing [38, 39].

Data Capture Setup Our goal while designing the capture setup was to find the best trade-off between maximizing data fidelity and interfering with the naturalness of the interaction (ecological validity) or violating participant privacy (ethical considerations). Through discussions with the HREC and General Chairs of MM'19 we decided to mitigate the capture of faces, which constitute one of the

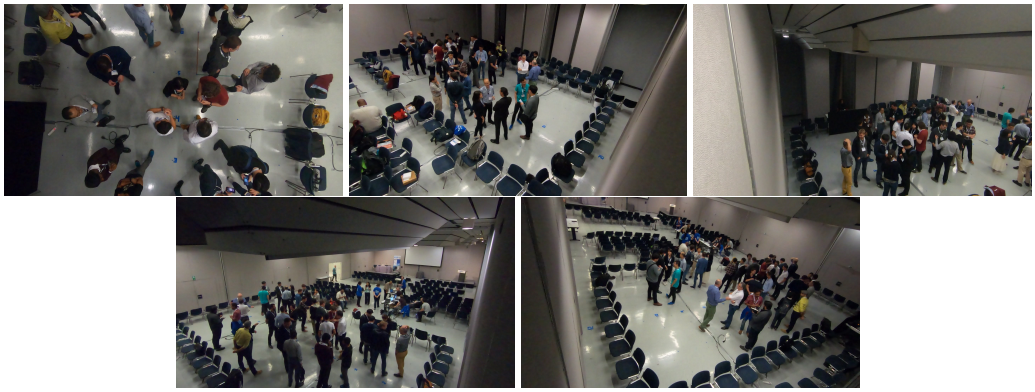


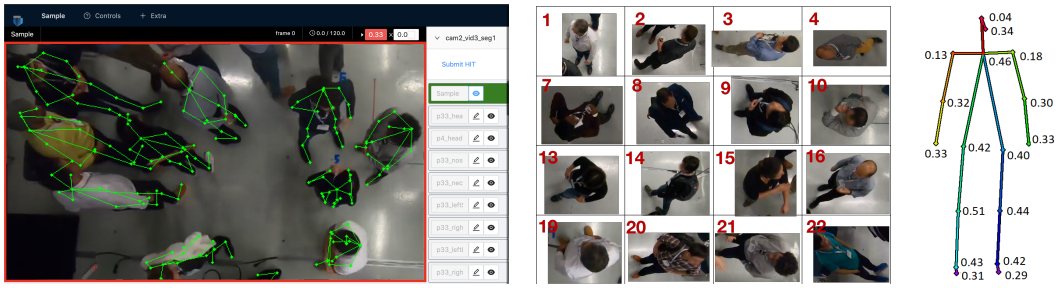
Figure 6: Comparing the top-down (top-left, camera 4) and elevated-side camera views (rest). Note how the top-down view is better at mitigating the capture of faces and suffers from fewer occlusions. This allows for a clearer capture of gestures and lower extremities for the most number of people while also preserving privacy.

most sensitive personally-identifiable features. Avoiding the inclusion of faces serves two purposes. First, it safeguards against misuse in downstream tasks with potential negative societal impacts such as harmful surveillance. Such issues have led to the retraction of some person re-identification datasets [14]. Second, it protects the participants who are part of a real research community; since the dataset does not involve role-playing or scripted conversations, the dataset contains their actual behavior. Consequently, we chose an aerial perspective for the video modality (see Figure 6). The $10\text{ m} \times 5\text{ m}$ interaction area was recorded by 14 GoPro Hero 7 Black cameras (60fps, 1080p, Linear, NTSC) [40]. 10 of these were placed directly overhead at a height of ~ 3.5 m at 1 m intervals, with 4 cameras at the corners providing an elevated-side-view perspective. (The HREC has suggested not sharing the elevated-side-view videos due to the presence of faces.) For capturing multimodal data streams, we designed a custom wearable multi-sensor pack called the Midge[2] (see Figure 5 for a design render), based on the open-source Rhythm Badge designed for office environments [41]. We improved upon the Rhythm Badge to achieve more fine-grained and flexible data capture (see Appendix D). We designed the Midge in a conference badge form-factor for seamless integration. Unlike smartphones, wearable badges allow for a simple *grab-and-go* setup and do not suffer from sensor/firmware differences across models. Popular human behavior datasets are synchronized by maximizing similarity scores around manually identified common events, such as infrared camera detections [11], or speech plosives [42]. While recordings in lab settings can allow for fully wired recording setups, recording in-the-wild requires a distributed wireless solution. We developed a solution to synchronize the cameras and wearable sensors directly at acquisition while significantly lowering the cost of the recording setup [18], making it easier for others to replicate our capture setup. See Appendix D for synchronization and calibration details, and Appendix B for images of the setup.

Data Association and Participant Protocol One consideration for multimodal data recording is the data association problem—how can pixels corresponding to an individual be linked to their other data streams? To this end, we designed a participant registration protocol. Arriving participants were greeted and fitted with a Midge. The ID of the Midge acted as the participant's identifier. One team member took a picture of the participant while ensuring both the face of the participant and the ID on the Midge were visible. In practice, it is preferable to avoid this step by using a fully automated multimodal association approach. However this remains an open research challenge [43, 44]. During the event, participants mingled freely—they were allowed to carry bags or use mobile phones. Conference volunteers helped to fetch drinks for participants. Participants could leave before the end of the one hour session.

Replicating Data Collection Setup and Community Engagement After the event, we gave a tutorial at MM'19 [45] to demonstrate how our collection setup could be replicated, and to invite conference attendees and event participants to reflect on the broader considerations surrounding privacy-preserving data capture, sharing, and future directions such initiatives could take.

[2]Documentation and schematics: https://github.com/TUDelft-SPC-Lab/spcl_midge_hardware



(a) Keypoint annotation interface in covfee [47] (b) Gallery of identities (c) Occlusion

Figure 7: Illustration of the body keypoints annotation procedure: (a): our custom time continuous annotation interface; (b): the gallery of person identities used by annotators to identify people in the scene (faces blurred); and (c): the skeleton template with the fraction of occluded frames.

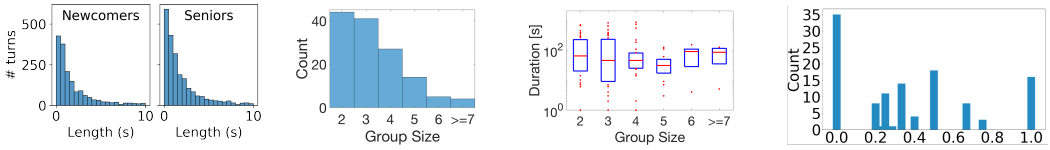
4 Data Annotation

Continuous Keypoints Annotation Existing datasets of in-the-wild social interactions have mainly focused on localizing subjects via bounding boxes [9, 11]. However, richer information about the social dynamics such as gestures and changes in orientation cannot be retrieved from bounding boxes alone, and necessitates the labeling of multiple skeletal keypoints. The typical approach to keypoint annotation involves using tools such as Vatic [25] or CVAT [46] to manually label every N frames followed by interpolating over the rest of the frames. This one-frame-at-a-time annotation procedure makes obtaining keypoint annotations a labor- and cost-intensive process. Moreover, interpolation fails to capture the finer temporal dynamics of the underlying behavior, and reduces the benefits of higher-framerate video capture. Limited by existing tools, no related dataset of in-the-wild human behavior has included time-continuous pose or speaking status annotations.

In contrast, to overcome these issues we collected fine-grained time-continuous annotations of keypoints via a web-based interface implemented as part of the Covfee framework [48]. Here, annotators follow individual joints using their mouse or trackpad while playing the video in their web browser. The playback speed of the video is automatically adjusted using an optical-flow-based technique to enable annotators to follow keypoints continuously without pausing the video. This design enables easy keypoint labeling in *every* frame of the video (60 Hz). We also incorporated a binary *occlusion* flag for every body keypoint. Annotators simultaneously controlled this flag to indicate when a body joint was not directly visible. Note that the flag is only an additional confidence indicator; we asked the annotators to label the occluded keypoint using their best estimate if it was deemed to be within the frame. Our pilot study on the efficacy of Covfee compared to non-continuous annotation via CVAT [46] is presented in [48]. For the pilot annotators, the continuous annotation methodology resulted in a $3\times$ speedup with statistically indifferent error rates.

We chose the top-down camera views for annotation since they suffer from fewer occlusions than the elevated-side views, enabling improved capture of gestures and lower extremities for more number of people (see Figure 6). Given the overlap in the camera views, we annotated keypoints in five of the ten overhead cameras (see Figure 1). Note that the same subject could be annotated in multiple cameras due to the overlap in even the five annotated cameras. Videos were split into two-minute segments to ease the annotation procedure. Each segment was annotated by one annotator by tracking the joints of all the people in the scene.

Continuous Speaking Status Annotations Speaking status is a key non-verbal cue for many social interaction analysis tasks [49]. We annotated the binary speaking status of every subject due to its importance as a key feature of social interaction [16, 50–53] and to contribute the existing community who are working on this task [17, 54, 55]. Action annotations have traditionally been carried out using frame-wise techniques [9], where annotators find the start and end frame of the action of interest using a graphical interface. Given the speed enhancement of continuous annotation, we also annotated speaking status via a continuous technique. We implemented a binary annotation interface as part of Covfee [48]. We asked annotators to press a key when they perceived speaking starting or ending. In a pilot study with two annotators, we measured a frame-level agreement (Fleiss' κ) of



(a) speaking turn lengths (b) group size (c) group duration (d) fraction of newcomers in groups

Figure 8: Data distributions for speaking status and conversation groups

0.552, comparable to previous work [35]. Similar to [9], the annotations were made by watching the video. We provided the annotators with all overhead views to best capture visual behavior.

F-formation Annotations Identifying who is likely to have social influence on whom is another important feature for analyzing social behavior. This is operationalised via the theory of F-formations, which are groups of people arranging themselves to converse or socially interact. Similar to prior datasets [9, 11, 13], F-formations group membership were annotated using an approximation of Kendon's definition [56]. F-formation stands for Facing formation, which is a socio-spatial arrangement where people have direct, easy and equal access while excluding the space from others in the surroundings. The arrangement commonly maintains a convex space in the middle of all the participants (determined by the location and orientation of their lower body), although other spatial arrangements (e.g., side-by-side, L-shaped) are possible, especially for smaller-sized groups of people. Annotations were labeled by one annotator at 1 Hz, following this definition. Since this is a largely objective and common framework for defining F-formations, we deemed it sufficient to obtain one set of annotations. Further, since F-formations may span camera views, we always used the camera that captured each F-formation in its entirety for annotation.

5 Dataset Statistics

Individual-Level Statistics Figure 7c shows the average occlusion values we obtained from annotators for each of the 17 keypoints. In Figure 8a we show the distribution of turn lengths in our speaking status annotations, for both newcomers and veterans, as per their self-reported newcomer status to the conference. We defined a turn to be a contiguous segment of positively-labeled speaking status, which resulted in a total of 4096 turns annotated.

Group-Level Statistics We found 119 distinct F-formations of size greater than or equal to two, and 38 instances of singletons. Of these, there are 14 F-formations and 2 singletons that include member(s) using the mobile phone. The distributions for group size and duration per group size are shown in Figure 8b and Figure 8c, respectively. Mean group duration doesn't seem to be influenced by group size although higher variations are seen at smaller group sizes. The fraction of community newcomers (first-time attending the conference) in groups is summarized in histogram in Figure 8d. The figure demonstrates two peaks on both sides of the spectrum (i.e., no newcomers vs. all newcomers in the same group). This spread over mixed and non-mixed seniority presents opportunities to study how acquaintance and seniority influence conversation dynamics.

6 Research Tasks

We report experimental results on three baseline benchmark tasks: person and keypoints detection, speaking status detection, and F-formation detection. The first task is a fundamental building block for automatically analyzing human social behaviors. The other two demonstrate how learned body keypoints can be used in the behavior analysis pipeline. We chose these benchmarking tasks since they have been commonly studied on other in-the-wild behavior datasets. Code for all benchmark tasks is available at: `https://github.com/TUDelft-SPC-Lab/conflab`. See the *Uses* section of the Datasheet in the Appendix for a discussion of the broader range of tasks ConfLab enables.

Table 2: Mask-RCNN results for person bounding box detection and keypoint estimation.

Model	Person Detection			Keypoint Estimation		
	AP_{50}	AP	AP_{75}	AP^{OKS}_{50}	AP^{OKS}	AP^{OKS}_{75}
R50-FPN	73.9	38.9	38.4	45.3	13.5	3.3

Figure 9: Predictions from the Mask-RCNN model; COCO pretrained (left), and ConfLab finetuned (right).

6.1 Person and Keypoints Detection

This benchmark involves the tasks of person detection (identifying bounding boxes) and pose estimation (localizing skeletal keypoints). Since pre-trained SOTA methods struggle with a privacy-sensitive top-down perspective [15] (also see Figure 3 and Appendix F.1 for ConfLab results), we finetune COCO-pretrained models on our dataset. We used Mask-RCNN [57] (Detectron2 framework [58] implementation) with a ResNet-50 backbone for both tasks for benchmarking. Since keypoint annotations were made per camera, we used four of the overhead cameras for training (Cameras 2, 4, 8, 10) and one for testing (Camera 6). Implementation details are available in Appendix E.1.

Evaluation Metrics We evaluated person-detection performance using the standard metrics in the MS-COCO dataset paper [59]. We report average precision (AP) for intersection over union (IoU) thresholds of 0.50 and 0.75, and the mean AP from an IoU range from 0.50 to 0.95 in 0.05 increments. For keypoint detection, we use object keypoint similarity (OKS) [59]. AP^{OKS} is a mean average precision for different OKS thresholds from 0.5 to 0.95.

Results and Analyses Table 2 summarizes our person detection and joint estimation results. Our baseline achieves 73.9 AP_{50} in detection and 45.3 AP^{OKS}_{50} in keypoint estimation. Figure 9 shows qualitative results from our fine-tuned network. For further insight we performed several analyses and ablations. In Appendix Table 6, we depict the effect of varying the number of training samples on performance. For training, we use the same four cameras and only vary the number of frames for each camera. We evaluate on the same testing images from camera 6. We find that performance saturates at 16% training samples. We next investigated the effect of increasing training data size by adding specific cameras one at a time. We report results in Appendix Table 7. There is a 260% performance gain when first doubling the training samples to 69 k with the addition of camera 4, and a 46% gain when adding another 43 k samples from camera 8. Finally, since the lower body regions suffer from higher occlusion, we experiment with different sections of body for further insight and report results in Appendix Table 8.

6.2 Speaking Status Detection

In data collected from real-life social settings, individual audio recordings can be hard to obtain due to privacy concerns [60]. This has led to the exploration of other modalities to capture some of the motion characteristics of speaking-related gestures [35, 36]. In this task we explore the use of body pose and wearable acceleration data for detecting the speaking status of a person in the scene.

Setup We use the SOTA MS-G3D graph neural network for skeleton action recognition [61], pre-trained on Kinetics Skeleton 400. For the acceleration modality, we evaluated three time series classifiers, each of which we trained from scratch: 1D Resnet [62], InceptionTime [63], and Minirocket [64]. We performed late fusion by averaging the scores from both modalities. Like prior work [17, 36], the task was set up as a binary classification problem. We divided our pose (skeleton) tracks into 3-second windows with 1.5 s overlap. A window was labeled positive if more than 50% of the continuous speaking status labels within it are positive. This resulted in an imbalanced dataset of 42882 windows with 29.2% positive labels. Poses were pre-processed for training following [61]. Three of the keypoints (head, and feet tips) were discarded due to not being present in Kinetics. We adapted the network by freezing all layers except for the last fully connected layer and training for five extra epochs. Acceleration readings were not pre-processed, other than by interpolating the original variable-sampling-rate signals to a fixed 50 Hz.

Table 3: ROC AUC and accuracy of skeleton-based, acceleration-based and multimodal speaking status detection (10-fold cross-validation).

Modality	Model	AUC	Acc.
Pose	MS-G3D [66]	0.676	0.677
	InceptionTime [63]	0.798	0.768
Acceleration	Resnet 1D [62]	0.801	0.767
	Minirocket [64]	0.813	0.768
Multimodal	MS-G3D + Minirocket	0.823	0.775

Table 4: Average F1 scores for F-formation detection comparing GTCG [23] and GCFF [67] with the effect of different threshold and orientations (standard deviation in parenthesis).

	GTCG		GCFF	
	T=2/3	T=1	T=2/3	T=1
Head	0.51 (0.09)	0.40 (0.12)	0.47 (0.07)	0.31 (0.23)
Shoulder	0.46 (0.11)	0.38 (0.11)	0.56 (0.25)	0.36 (0.16)
Hip	0.45 (0.10)	0.37 (0.12)	0.39 (0.06)	0.25 (0.11)

Evaluation Evaluation was carried out via 10-fold cross-validation at the subject level, ensuring that no examples from the test subjects were used in training. We used the area under the ROC curve (AUC) as main evaluation metric to account for the imbalance in the labels.

Results The results in Table 3 indicate a better performance from the acceleration-based methods. One possible reason for the lower performance of the pose-based methods is the significant domain shift between Kinetics and Conflab, especially in camera viewpoint (frontal vs top-down). The acceleration performance is in line with previous work [17]. Multimodal results were slightly higher than acceleration-only results, despite our naive fusion approach, a possible point to improve in future work [65]. Experiments with the rest of the IMU modalities are presented in Appendix F.2.

6.3 F-formation Detection

Setup Like prior work [10, 21–23], we operationalize interaction groups using the framework of F-formations [56]. We provide performance results for F-formation detection using GTCG [23] and GCFF [67] as a baseline. Recent deep learning methods such as DANTE [22] are not directly applicable since they depend on knowing the number of people in the scene, which is variable for ConfLab. We use pre-trained model parameters (reported in the original GTCG and GCFF papers on the Cocktail Party dataset [13]) and tuned a subset of parameters more relevant to ConfLab attributes on camera 6. More details can be found in Appendix E.2. We derive three different sets of orientation features from (i) head, (ii) shoulder and (iii) hip keypoints.

Evaluation Metrics We use the standard F1 score as evaluation metric for group detection [23, 67]. A group is correctly estimated (true positive) if at least $\lceil T * |G| \rceil$ of the members of group G are correctly identified, and no more than $1 - \lceil T * |G| \rceil$ is incorrectly identified, where T is the tolerance threshold. We report results for $T = \frac{2}{3}$ and $T = 1$ (more strict threshold) in Table 4.

Results We show that different results are obtained using different sources of orientations. Different occlusion levels in keypoints due to camera viewpoint may have affected performance. Another factor influencing model performance is that F-formations (which are driven by lower-body orientations [56]) may have multiple conversations floors [51]. Floors are indicated by coordinated speaker turn taking patterns and influence coordinated head orientations of the group.

7 Conclusion and Discussion

ConfLab contributes a new concept for real-life data collection in the wild and captures a high-fidelity dataset of mixed levels of acquaintance, seniority, and personal motivations.

ConfLab: the Dataset We improved upon prior work by providing higher-resolution, fidelity, and synchronization across sensor networks. We also carefully designed our social interaction setup to enable a diverse mix of seniority, acquaintanceship, and motivations for mingling. The result is a rich set of 17 body-keypoint annotations of 48 people at 60 Hz from overhead cameras for developing more robust estimation of keypoints, speaking status and F-formations for further analyses of more complex socio-relational phenomena. Our benchmark results for these tasks highlight how the improved fidelity of ConfLab can assist in the development of more robust methods for these key tasks. We hope that models trained on ConfLab for localizing keypoints would fill the gap in the cue

extraction pipeline, enabling past datasets [9, 10] without articulated pose data to be reinvigorated; this would open the floodgates for more robust analysis of the social phenomena labeled in these other datasets. Finally, our baseline social tasks form the basis for further explorations into downstream prediction tasks of socially-related constructs such as conversation quality [68] , dominance [53], rapport [50], influence [69] etc.

ConfLab: the Data-Collection Concept To relate an individual's behaviors to trends within their social network, further iterations of ConfLab are needed. These iterations would enable the study of behavioral patterns at different timescales, including multiple interactions in one day, multiple days at a conference, or across distinct conferences. This paper serves as a template for such future ventures. We hope that if the idea of a conference as a living lab gains traction, the effort and cost of data collection can be amortized across different research groups, even involving support from the conference organizers. This *data by the community for the community* ethos can enable the generation of a corpus of related datasets enabling new research questions.

Societal Impact ConfLab's long-term vision is towards developing technology to assist individuals in navigating social interactions. In this work we have identified choices that maximize data fidelity while upholding ethical best practices: an overhead camera perspective that mitigates identifying faces, recording audio at a low-frequency, and using non-intrusive wearable sensors matching a conference badge form-factor. We argue this is an essential step towards a long-term goal of developing personalized and socially aware technologies that enhance social experiences. At the same time, such interventions could also affect a community in unintended ways: worsened social satisfaction, lack of agency, stereotyping; or benefit only those members of the community who make use of resulting applications at the expense of the rest. More nefarious uses involve exploiting the data for developing methods that harmfully surveil or profile people. Researchers must consider such inadvertent effects while developing downstream applications. Finally, since we recorded the dataset at a scientific conference and required voluntary participation, there is an implicit selection bias in the population represented in the data. Researchers should be aware that insights resulting from the data may not generalize to the general population.

Empowering Users Through an Agentist Rather Than Structurist Approach The analysis of human behavior in social settings has classically taken a more top-down perspective. For instance, the analysis of situated interactions (via only proximity networks) has provided insight into the process of making science in the field of Meta Science [70]. However, while social network science is a well-populated domain, it lacks a more individualized measurement of social behavior: see more discussion of the structure vs. agency debate [71]. Relying on the network science approach jeopardizes an individual's right to technologies that enable free will. We consider the agency in choosing such technologies to be a form of individual harm avoidance. ConfLab provides access to more than just proximity data about social interactions, enabling the study of context-specific social dynamics. These dynamics are a uniquely dependent not only on the individual, but also the group they are interacting with [72]. We hope our highlighting of participatory design practices and these value-sensitive design principles promote social safety in developing socially assistive technologies.

Acknowledgements

The authors would like to thank: the ACM Multimedia 2019 General Chairs Martha Larson, Benoit Huet, and Laurent Amsaleg for their support in making the data collection at a major international conference a reality; Bernd Dudzik, Yeshwanth Napolean, Ruud de Jong, and the venue support staff for their help in setting up the recording on site; Ioannis Protonotarios for the development of the MINGLE Midge badge; Jerry de Vos for improving our Midge Github repository and designing a new case; the participants and student volunteers for the *Meet the Chairs!* event; the Amazon Mechanical Turk workers for their efforts in annotating the dataset; Rich Radke, Martin Atzmueller, Laura Cabrera-Quiros, Alan Hanjalic, and Xucong Zhang for the insightful discussions; Santosh Ilamparuthi for the innumerable discussions and support towards strengthening the ethical soundness of recording and sharing ConfLab; Jan van der Heul for the incredibly responsive support in setting up the 4TU Data repository for ConfLab; and Bart Vastenhouw, Myrthe Tielman, and Catharine Oertel for help with the data sharing; and Musy Ayoub for the word-intelligibility analysis of the low frequency audio.

ConfLab was partially funded by Netherlands Organization for Scientific Research (NWO) under project number 639.022.606 with associated Aspasia Grant, and also by the ACM Multimedia 2019 conference via student helpers, and crane hiring for camera mounting.

References

[1] Bernd Dudzik, Simon Columbus, Tiffany Matej Hrkalovic, Daniel Balliet, and Hayley Hung. Recognizing perceived interdependence in face-to-face negotiations through multimodal analysis of nonverbal behavior. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 121–130, 2021. 1

[2] William Fleeson. Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of personality*, 75:825–862, 8 2007. ISSN 0022-3506. doi: 10.1111/J.1467-6494.2007.00458.X. URL https://pubmed.ncbi.nlm.nih.gov/17576360/. 1

[3] Jennifer G. La Guardia and Richard M. Ryan. Why identities fluctuate: Variability in traits as a function of situational variations in autonomy support. *Journal of Personality*, 75:1205–1228, 12 2007. ISSN 00223506. doi: 10.1111/j.1467-6494.2007.00473.x.

[4] Judith A. Hall, Terrence G. Horgan, and Nora A. Murphy. Nonverbal communication. *https://doi.org/10.1146/annurev-psych-010418-103145*, 70:271–294, 1 2019. ISSN 15452085. doi: 10.1146/ANNUREV-PSYCH-010418-103145. URL https://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010418-103145. 1

[5] Katherine Osborne-Crowley. Social cognition in the real world: reconnecting the study of social cognition with social reality. *Review of general psychology*, 24(2):144–158, 2020. 1

[6] Chittaranjan Andrade. Internal, external, and ecological validity in research design, conduct, and evaluation. *Indian journal of psychological medicine*, 40(5):498–499, 2018. 2, 4

[7] Élise Labonte-LeMoyne, François Courtemanche, Marc Fredette, and Pierre-Majorique Léger. How wild is too wild: Lessons learned and recommendations for ecological validity in physiological computing research. In *PhyCS*, pages 123–130, 2018.

[8] Hayley Hung, Ekin Gedik, and Laura Cabrera Quiros. Complex conversational scene analysis using wearable sensors. In *Multimodal Behavior Analysis in the Wild*, pages 225–245. Elsevier, 2019. 2, 4

[9] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. The matchnmingle dataset: A novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, 12(1):113–130, 2021. 2, 3, 4, 6, 7, 10, 12

[10] Hayley Hung and Ben Kröse. Detecting f-formations as dominant sets. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 231–238, 2011. 2, 3, 9, 10, 12

[11] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1707–1720, 2015. 2, 3, 4, 5, 6, 7, 12

[12] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. Social interaction discovery by statistical analysis of f-formations. In Jesse Hoey, Stephen J. McKenna, and Emanuele Trucco, editors, *British Machine Vision Conference, BMVC 2011, Dundee, UK, August 29 - September 2, 2011. Proceedings*, pages 1–12. BMVA Press, 2011. doi: 10.5244/C.25.23. URL https://doi.org/10.5244/C.25.23. 3

[13] Gloria Zen, Bruno Lepri, Elisa Ricci, and Oswald Lanz. Space speaks: towards socially and personality aware visual surveillance. In *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*, pages 37–42, 2010. 2, 3, 7, 9

[14] Madhumita Murgia. Who's using your face? the ugly truth about facial recognition. *Financial Times*, 2019. 2, 5

[15] Nicolo Carissimi, Paolo Rota, Cigdem Beyan, and Vittorio Murino. Filling the gaps: Predicting missing joints of human poses using denoising autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2, 8

[16] Ekin Gedik and Hayley Hung. Detecting conversing groups using social dynamics from wearable acceleration: Group size awareness. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(4), dec 2018. doi: 10.1145/3287041. URL https://doi.org/10.1145/3287041. 2, 6

[17] Ekin Gedik and Hayley Hung. Personalised models for speech detection from body movements using transductive parameter transfer. *Personal and Ubiquitous Computing*, 21(4):723–737, August 2017. ISSN 1617-4909. doi: 10.1007/s00779-017-1006-4. 2, 3, 4, 6, 8, 9, 12

[18] Chirag Raman, Stephanie Tan, and Hayley Hung. A modular approach for synchronized wireless multimodal multisensor data acquisition in highly dynamic social settings. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 3586–3594, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413697. URL https://doi.org/10.1145/3394171.3413697. 2, 3, 4, 5, 8, 12, 16

[19] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xiangyu Zhang, Xinyu Zhou, Erjin Zhou, and Jian Sun. Learning delicate local representations for multi-person pose estimation. In *European Conference on Computer Vision*, pages 455–472. Springer, 2020. 2, 17, 18

[20] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 1536–1546. IEEE, 2021. doi: 10.1109/WACV48630.2021.00158. URL https://doi.org/10.1109/WACV48630.2021.00158. 2, 3

[21] Francesco Setti, Chris Russell, Chiara Bassetti, and Marco Cristani. F-formation detection: Individuating free-standing conversational groups in images. *PloS one*, 10(5):e0123783, 2015. 3, 9

[22] Mason Swofford, John Peruzzi, Nathan Tsoi, Sydney Thompson, Roberto Martín-Martín, Silvio Savarese, and Marynel Vázquez. Improving social awareness through dante: Deep affinity network for clustering conversational interactants. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23, 2020. 9

[23] Sebastiano Vascon, Eyasu Zemene Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. A game-theoretic probabilistic approach for detecting conversational groups. In *Asian conference on computer vision*, pages 658–675. Springer, 2014. 3, 9

[24] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3, 12

[25] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently Scaling up Crowdsourced Video Annotation. *International Journal of Computer Vision*, 101(1):184–204, 2013. ISSN 0920-5691. doi: 10.1007/s11263-012-0564-1. 3, 6

[26] Elisa Ricci, Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulò, Narendra Ahuja, and Oswald Lanz. Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4660–4668, 2015. 3

[27] Loris Bazzani, Marco Cristani, Diego Tosato, Michela Farenzena, Giulia Paggetti, Gloria Menegaz, and Vittorio Murino. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 30(2):115–127, 2013. 3

[28] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The ami meeting corpus: A pre-announcement. In Steve Renals and Samy Bengio, editors, *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. 3

[29] C. Cattuto, W. V. D. Broeck, A. Barrat, V. Colizza, J. Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLoS ONE*, 5, 2010. 3

[30] Marion Hoffman, Per Block, Timon Elmer, and Christoph Stadtfeld. A model for the dynamics of face-to-face interactions in social groups. *Network Science*, 8(S1):S4–S25, 2020. doi: 10.1017/nws.2020.3.

[31] Martin Atzmueller and Florian Lemmerich. Homophily at academic conferences. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 109–110, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[32] Daniel Olguín Olguín, Benjamin N Waber, Taemie Kim, Akshay Mohan, Koji Ara, and Alex Pentland. Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):43–55, 2008. 3

[33] Daniel Chaffin, Ralph Heidl, John R Hollenbeck, Michael Howe, Andrew Yu, Clay Voorhees, and Roger Calantone. The promise and perils of wearable sensors in organizational research. *Organizational Research Methods*, 20(1):3–31, 2017. 3

[34] Alessio Rosatelli, Ekin Gedik, and Hayley Hung. Detecting f-formations roles in crowded social scenes with wearables: Combining proxemics dynamics using lstms. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 147–153, 2019. doi: 10.1109/ACIIW.2019.8925179. 4

[35] Laura Cabrera-Quiros, David M.J. Tax, and Hayley Hung. Gestures in-the-wild : Detecting conversational hand gestures in crowded scenes using a multimodal fusion of bags of video trajectories and body worn acceleration. pages 1–10, 2018. 4, 7, 8, 3, 5, 12

[36] J. V. Quiros and H. Hung. CNNs and Fisher Vectors for No-Audio Multimodal Speech Detection. In *MediaEval*, 2019. 4, 8, 3, 12

[37] Stephanie Tan, David M. J. Tax, and Hayley Hung. Multimodal joint head orientation estimation in interacting groups via proxemics and interaction dynamics. *Proc. ACM Interactive, Mobile, Wearable, and Ubiquitous Technology*, 5(1), March 2021. 4

[38] University of york research data management. https://www.york.ac.uk/library/info-for/researchers/data/sharing/access/. 4

[39] Utrecht university research data management. https://www.uu.nl/en/research/research-data-management/guides/handling-personal-data. 4

[40] Go pro hero 7 black. https://gopro.com/en/nl/shop/cameras/hero7-black/CHDHX-701-master.html. 5

[41] Oren Lederman, Akshay Mohan, Dan Calacci, and Alex Sandy Pentland. Rhythm: A unified measurement platform for human organizations. *IEEE MultiMedia*, 25(1):26–38, 2018. 5

[42] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013. 5

[43] Laura Cabrera-Quiros and Hayley Hung. Who is where? matching people in video to wearable acceleration during crowded mingling events. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 267–271, 2016. 5, 12

[44] Laura Cabrera-Quiros and Hayley Hung. A hierarchical approach for associating body-worn sensors to video regions in crowded mingling scenarios. *IEEE Transactions on Multimedia*, 21(7):1867–1879, 2018. 5, 12

[45] Hayley Hung, Chirag Raman, Ekin Gedik, Stephanie Tan, and Jose Vargas Quiros. Multimodal data collection for social interaction analysis in-the-wild. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2714–2715, 2019. 5

[46] Computer Vision Annotation Tool (CVAT). 6

[47] Covfee: Continuous Video Feedback Tool. Jose Vargas. 6

[48] Jose Vargas Quiros, Stephanie Tan, Chirag Raman, Laura Cabrera-Quiros, and Hayley Hung. Covfee: an extensible web framework for continuous-time annotation of human behavior. In Cristina Palmero, Julio C. S. Jacques Junior, Albert Clapés, Isabelle Guyon, Wei-Wei Tu, Thomas B. Moeslund, and Sergio Escalera, editors, *Understanding Social Behavior in Dyadic and Small Group Interactions*, volume 173 of *Proceedings of Machine Learning Research*, pages 265–293. PMLR, 16 Oct 2022. URL https://proceedings.mlr.press/v173/vargas-quiros22a.html. 6, 5, 8

[49] Daniel Gatica-Perez. Analyzing group interactions in conversations: a review. In *2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 41–46, 2006. doi: 10.1109/MFI.2006.265658. 6

[50] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In *23rd International Conference on Intelligent User Interfaces*. ACM, 2018. ISBN 978-1-4503-4945-1. doi: 10.1145/3172944.3172969. 6, 10, 12

[51] Chirag Raman and Hayley Hung. Towards automatic estimation of conversation floors within F-formations. *arXiv:1907.10384 [cs]*, July 2019. 9, 12

[52] Hayley Hung and Daniel Gatica-Perez. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia*, 12(6):563–575, 2010.

[53] Hayley Hung, Yan Huang, Gerald Friedland, and Daniel Gatica-Perez. Estimating Dominance in Multi-Party Meetings Using Speaker Diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):847–860, may 2011. 6, 10, 12

[54] Cigdem Beyan, Muhammad Shahid, and Vittorio Murino. RealVAD: A Real-world Dataset and A Method for Voice Activity Detection by Body Motion Analysis. *x*, 9210(c):1–16, 2020. doi: 10.1109/tmm.2020.3007350. 6

[55] Muhammad Shahid, Cigdem Beyan, and Vittorio Murino. Voice activity detection by upper body motion analysis and unsupervised domain adaptation. *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*, pages 1260–1269, 2019. doi: 10.1109/ICCVW.2019.00159. 6

[56] Adam Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990. 7, 9

[57] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 8

[58] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 8

[59] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 8

[60] Jiaxing Shen, Oren Lederman, Jiannong Cao, Florian Berg, Shaojie Tang, and Alex Sandy Pentland. GINA: Group Gender Identification Using Privacy-Sensitive Audio Data. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2018-Novem:457–466, 2018. ISSN 15504786. doi: 10.1109/ICDM.2018.00061. 8

[61] Pranay Gupta, Anirudh Thatipelli, Aditya Aggarwal, Shubh Maheshwari, Neel Trivedi, Sourav Das, and Ravi Kiran Sarvadevabhatla. Quo Vadis, Skeleton Action Recognition ? *arXiv:2007.02072 [cs]*, July 2020. 8, 19

[62] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline, 2016. URL https://arxiv.org/abs/1611.06455. 8, 9

[63] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. *Proceedings of the International Joint Conference on Neural Networks*, 2017-May:1578–1585, 2017. doi: 10.1109/IJCNN.2017.7966039. 8, 9

[64] Chang Wei Tan, Angus Dempster, Christoph Bergmeir, and Geoffrey I. Webb. Multirocket: Multiple pooling operators and transformations for fast and effective time series classification, 2021. URL https://arxiv.org/abs/2102.00457. 8, 9, 17, 19

[65] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, February 2019. 9, 13

[66] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, October 2020. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2019.2916873. 9

[67] Marco Cristani, Ramachandra Raghavendra, Alessio Del Bue, and Vittorio Murino. Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, 100:86–97, 2013. 9

[68] Chirag Raman, Navin Raj Prabhu, and Hayley Hung. Perceived conversation quality in spontaneous interactions, 2022. URL https://arxiv.org/abs/2207.05791. 10, 12

[69] Wen Dong, Bruno Lepri, Alessandro Cappelletti, Alex Sandy Pentland, Fabio Pianesi, and Massimo Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 271–278, 2007. 10, 12

[70] Julia Eberle, Karsten Stegmann, Alain Barrat, Frank Fischer, and Kristine Lund. Initiating scientific collaborations across career levels and disciplines–a network analysis on behavioral data. *International Journal of Computer-Supported Collaborative Learning*, 16(2):151–184, 2021. 10

[71] Nigel Pleasants. Free will, determinism and the "problem" of structure and agency in the social sciences. *Philosophy of the Social Sciences*, 49(1):3–30, 2019. 10

[72] Chirag Raman, Hayley Hung, and Marco Loog. Social Processes: Self-Supervised Meta-Learning over Conversational Groups for Forecasting Nonverbal Social Cues. *arXiv:2107.13576 [cs]*, July 2021. 10, 12

[73] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021. 1, 3

[74] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research: a report from the neurips 2019 reproducibility program. *Journal of Machine Learning Research*, 22, 2021. 1

[75] German Barquero, Johnny Núnez, Sergio Escalera, Zhen Xu, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero. Didn't see that coming: a survey on non-verbal social human behavior forecasting. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 139–178. PMLR, 2022. 12

[76] Chirag Raman, Hayley Hung, and Marco Loog. Why did this model forecast this future? closed-form temporal saliency towards causal explanations of probabilistic forecasts. *arXiv preprint arXiv:2206.00679*, 2022. 12

[77] Hayley Hung, Gwenn Englebienne, and Jeroen Kools. Classifying social actions with a single accelerometer. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 207–210, 2013. 12

[78] Navin Raj Prabhu, Chirag Raman, and Hayley Hung. Defining and quantifying conversation quality in spontaneous interactions. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 196–205, 2020. 12

[79] Jose David Vargas Quiros, Oyku Kapcak, Hayley Hung, and Laura Cabrera-Quiros. Individual and joint body movement assessed by wearable sensing as a predictor of attraction in speed dates. *IEEE Transactions on Affective Computing*, 2021. 12

[80] OpenCV. Open source computer vision library. https://github.com/opencv/opencv, 2015. 16

[81] Idiap multi camera calibration suite. https://github.com/idiap/multicamera-calibration. 16

[82] Tdkicm20948. https://invensense.tdk.com/products/motion-tracking/9-axis/icm-20948/. Accessed: 2021-10-15. 16

[83] Sileye O Ba and Jean-Marc Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):101–116, 2010. 17

[84] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019. 17, 18

[85] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 17, 18

[86] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017. 17, 18

[87] Ignacio Oguiza. tsai - a state-of-the-art deep learning library for time series and sequential data. Github, 2022. URL https://github.com/timeseriesAI/tsai. 19