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

Appendices

The Appendices are organized as follows:

- Section A: Hosting, licensing, and organization information for ConfLab
- Section B: Documentation for ConfLab, following Datasheets for Datasets [73]
- Section C: Sample post-hoc behavioral analysis report sent to each ConfLab participant
- Section D: Details about out data-capture setup
- Section E: Implementation details for models used in our benchmark research tasks
- Section F: Additional experimental results and ablations
- Section G: Details for reproducibility following the ML Reproducibility Checklist [74]

A Hosting, Licensing, and Organization

The dataset is hosted by 4TU.ResearchData, available at https://doi.org/10.4121/c.6034313.

The dataset itself is available under restricted access defined by an End-User License Agreement (EULA). The EULA itself is available under a CC0 license. The code (https://github.com/TUDelft-SPC-Lab/conflab) for the benchmark baseline tasks, and the schematics and data associated with the design of our custom wearable sensor called the Midge (https://github.com/TUDelft-SPC-Lab/spcl_midge_hardware) are available under the MIT License.

Figure 10 on the next page illustrates the organization of the ConfLab dataset on 4TU.ResearchData. The components are as follows:

- Annotations (restricted, https://doi.org/10.4121/20017664):
  annotations of pose, speaking status, and F-formations
- Datasheet for ConfLab (public, https://doi.org/10.4121/20017559):
  documentation of the dataset following Datasheets for Datasets [73] (see Appendix B)
- EULA (public, https://doi.org/10.4121/20016194):
  End User License Agreement to be signed for requesting access to the restricted components
- Processed-Data (restricted, https://doi.org/10.4121/20017805):
  processed video and wearable sensor used for annotations
- Raw-Data (restricted, https://doi.org/10.4121/20017748):
  raw video and wearable sensor data
- Data Samples (restricted, https://doi.org/10.4121/20017682):
  samples of the sensor, audio, and video data

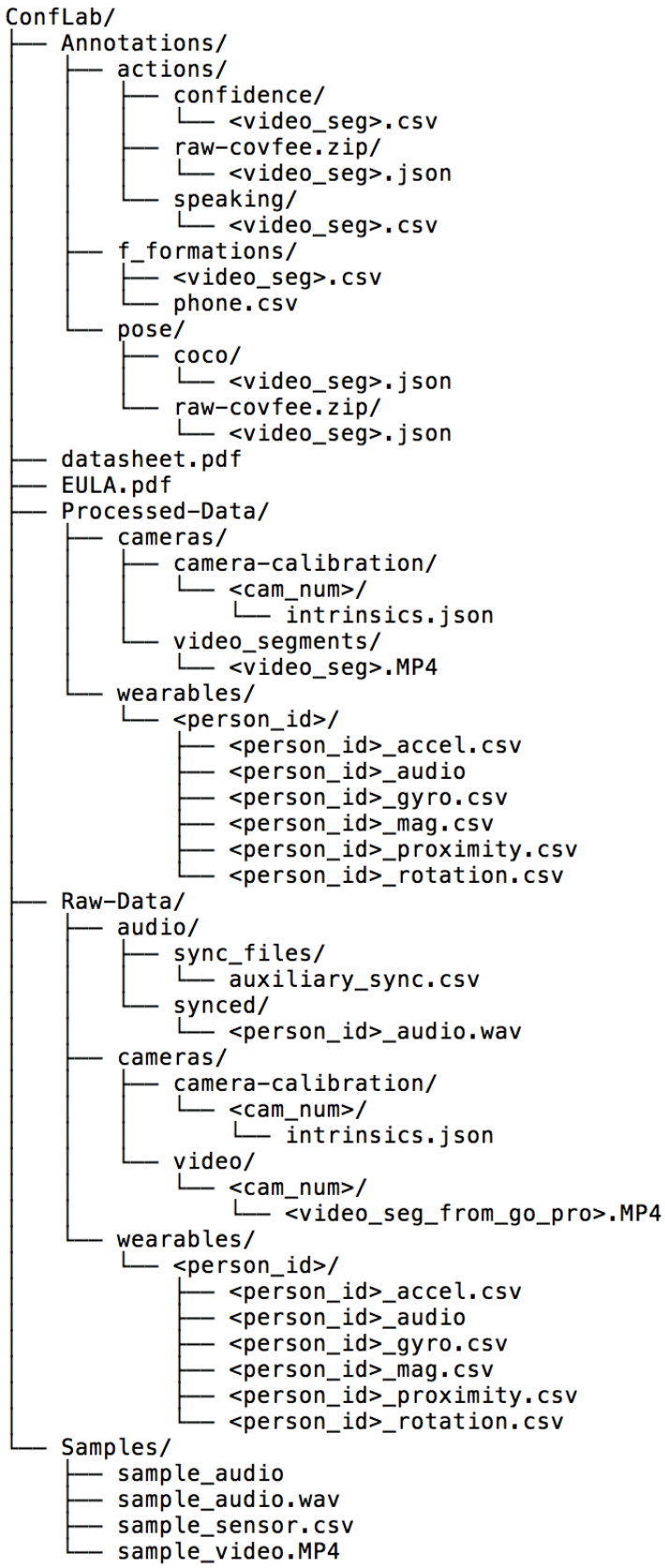


Figure 10: File structure of the ConfLab dataset

B Datasheet For ConfLab

This document is based on *Datasheets for Datasets* by Gebru *et al.* [73]. Please see the most updated version here.

MOTIVATION

Q. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

There are two broad motivations for creating this dataset: first, to enable the privacy-preserving, multimodal study of *real-life* social conversation dynamics; second, to bring the higher fidelity of wired in-the-lab recording setups to in-the-wild scenarios, enabling the study of *fine time-scale* social dynamics in-the-wild.

We propose the Conference Living Lab (ConfLab) with the following goals: (i) a data collection effort that follows 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 [20], we incorporated privacy and invasiveness considerations directly into the decision-making process regarding sensor type, positioning, and sample-rates.

From a technical perspective, closest related datasets (see Table 1 in the main paper) 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]. This often requires modeling simplifications such as the summarizing of features over rolling windows [17, 35, 36]. On the other hand, past high-fidelity datasets have largely involved role-played or scripted interactions in lab settings, with often a single-group in the scene.

This dataset wasn't created with a specific task in mind, but intends to support a wide variety of multimodal modeling and analysis tasks across research domains (see the *Uses* section).

Q. Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

ConfLab was initiated by the Socially Perceptive Computing Lab, Delft University of Technology in cooperation and support from the general chairs of ACM Multimedia 2019 (Martha Larson, Benoit Huet, and Laurent Amsaleg), Nice, France. Since this dataset was by the community, for the community, members of the Multimedia community contributed as subjects in the dataset.

Q. What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

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.

Q. Any other comments?
None.

COMPOSITION

Q. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The dataset contains multimodal recordings of people interacting during a networking event embedded in an international multimodal machine learning conference.

Overall, the interaction scene contained conversation groups (operationalized as f-formations), composed of individual subjects, each of which had individual data associated to their wearable sensors. The complete interaction scene was additionally captured by overhead cameras. Figure 11 shows the structure of these instances and their relationships.

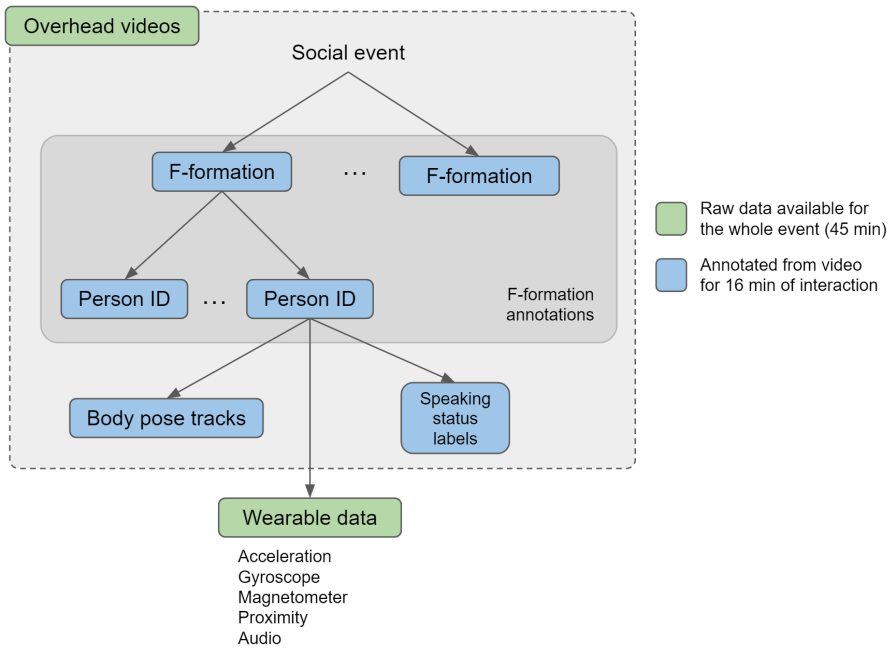


Figure 11: Structure of some of the instances in the dataset and their relationships. The interaction space was captured via overhead videos, in which f-formations (conversation groups) were annotated. An F-formation consists of set of people interacting for a variable period of time, and identified via a subject ID. Each person in the F-formation can be associated to their pose (annotated in the videos), their wearable sensor (IMU) data, and their action (speaking status) labels.

Note however that the precise notion of what constitutes an instance in the dataset is very much task-specific. In our baseline tasks we considered the following instances:

Person and Keypoints Detection Frames, containing pose annotations (17 body keypoints per person per frame @60 Hz) from 5 overhead videos (1920×1080, 60 fps) for 16 minutes of interaction.

Speaking Status Detection Windows (3 seconds) of wearable sensor data and speaking status annotations (60 Hz) extracted from each subject's data.

F-formations Operationalized conversation groups, annotated at 1 Hz from the 16 minutes of annotated data, and the pose data associated to the people in the F-formation.

Q. How many instances are there in total (of each type, if appropriate)?
The notion of instance is very much dependent on how a user intends to use the data. Regarding the instances in Figure 11, our full dataset consist of 45 minutes of:

Video recordings from 10 overhead cameras placed over the interaction area. Five of these videos, enough to cover the complete interaction area, were used in annotation.

Individual wearable sensor data For the 48 subjects in the interaction area, a chest-worn conference-type badge recorded: audio (1250 Hz), and Inertial Measurement Unit (IMU) readings (accelerometer @ 56 Hz, gyroscope @56 Hz, magnetometer @56 Hz and Bluetooth RSSI-based proximity @5 Hz)

Conference experience label For each of the 48 subjects, an associated self-report label indicating whether it was their first time in the conference.

The instances in the annotated 16 minutes segment out of the 45 minutes of interaction contain:

2D body poses For each of the 48 subjects, full body pose tracks annotated at 60Hz (17 keypoints per person). These were annotated using 5 of the 10 overhead cameras due to the significant overlap in views (cameras 2, 4, 6, 8, and 10). Annotations were done separately for each camera by annotating all of the people visible in each video, for each of the 5 cameras, and tagged with a participant ID. We made use of a novel continuous technique for annotation of keypoints. We chose this approach via a pilot study with 3 annotators, comparing our technique to annotations done using the non-continuous CVAT tool. We found no statistically significant differences in errors per-frame (as measured using Mean Squared Error across annotators), despite a 3x speed-up in annotation time in the continuous condition. The details of the technique and this pilot study can be found in [48].

Speaking status annotations For each of the 48 subjects, these include a) a binary signal (60 Hz) indicating whether the person is perceived to be speaking or not; b) continuous confidence value (60 Hz) indicating the degree of confidence of the annotator in their speaking status assessment. These annotations were done without access to audio due to issues with the synchronization of the audio recordings at the time of annotation. The confidence assessment is therefore largely based on the visibility of the target person and their speaking-associated gestures (eg. occlusion, orientation w.r.t. camera, visibility of the face)? We measured inter-annotator agreement for speaking status in a pilot where two annotators labeled three data subjects for 2 minutes each. We measured a frame-level agreement (Fleiss' κ) of 0.552, comparable to previous work [35].

F-formation annotations These annotations label the conversing groups in the scene following previous work. Each individual belongs to one F-formation at a time or is a singleton in the interaction scene. The membership is binary. The annotations were done by one of the authors at 1 Hz by watching the video. The time-stamped usage of mobile phones are available as auxiliary annotations, which are useful for the study of the role of mobile phone users as associates of F-formations. Since Kendon's theories date back to before the widespread use of mobile phones, their influence on F-formation membership remains an open question.

In our baseline tasks, which made use of the complete annotated section of the dataset, the instance numbers were the following:

Person and Keypoints Detection 119k frames (60fps) containing 1967k person instances (poses) in total, from 48 subjects recorded in 5 cameras (16 minutes of annotated segment).

Speaking Status Detection 42884 3-second windows, extracted from the 48 participants' wearable data and speaking status annotations.

F-formations 119 conversation groups. Details are in Section 5.

Q. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The participants in our data collection are a sample of the conference attendees. Participants were recruited via the conference website, social media posting, and approaching them in person during the conference. Because participation in such a data collection can only be voluntary, the sample was not pre-designed and may not be representative of the larger set.

Additionally, 16 minutes of sensor data has been annotated for keypoints, speaking status and F-formations out of the total of 45 minutes recorded. The remaining part (across all modalities) is provided with no labels. For privacy reasons, the elevated cameras (distinct from the previously mentioned 8 overhead cameras) and also individual frontal headshots that were used for manually associating the video data to the wearable sensor data is not being shared.

Q. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Camera 5 failed early during the recording, but the space underneath it was captured by the adjacent cameras due to the high overlap in the camera field-of-views. Nevertheless we share what was recorded before the failure from camera 5, bringing the total number of cameras to 9.

Q. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

The F-formations, subjects, and their associated data relate as shown in Figure 11. These associations are made explicit in the dataset via anonymous subject IDs, associated to pose tracks, speaking status annotations, and wearable sensor data. These same IDs were used to annotate the F-formations.

Pre-existing personal relationships between the subjects were not requested for privacy reasons.

Q. Are there recommended data splits (e.g., training, development/validation, testing)?

Since the dataset can be used to study a variety of tasks, the answer to this question is task dependent. Please refer to our reproducibility details (Appendix G of our associated paper) for information about the splits that we used in out baselines.

Q. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Individual audio Because audio was recorded by a front-facing wearable device worn on the chest, it contains a significant amount of cocktail party noise and cross-contamination from other people in the scene. In our experience this means that automatic speaking status detection is challenging with existing algorithms but manual annotation is possible.

Videos and 2D body poses It is important to consider that the same person may appear in multiple videos at the same time if the person was in view of multiple cameras. Because 2D poses were annotated per video, the same is true of pose annotations. Each skeleton was tagged with a person ID, which should serve to identify such cases when necessary.

Q. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained.

Q. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?

The data contains personal data under GDPR in the form of video and audio recordings of subjects. The dataset is shared under an End User License Agreement for research purposes, to ensure that the data is not made public, and to protect the privacy of data subjects.

Q. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No.

Q. Does the dataset relate to people?

Yes, the dataset contains recordings of human subjects.

Q. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Data subjects answered the following questions before the start of the data collection event, after filling in their consent form:

- Is this your first time attending ACM MM?
- Select the area(s) that describes best your research interest(s) in recent years. Descriptions of each theme are listed here: https://acmmm.org/call-for-papers/

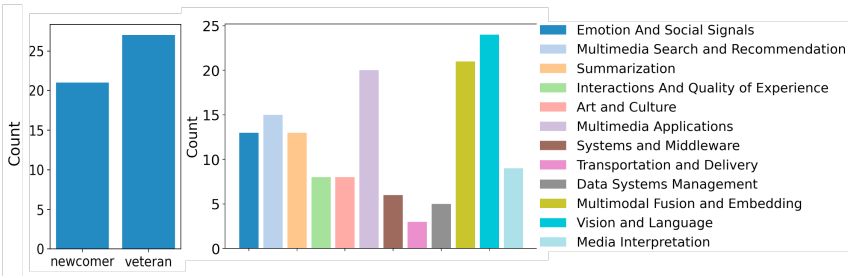


Figure 12: Distribution of participant seniority (left) and research interests (right) in percentage.

Figure 12 shows the distribution of the responses / populations.

Q. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

We do not share any directly identifiable information as part of the dataset. However, individuals may be identified in the video recordings if the observer knows the participants in the recordings personally. Otherwise, individuals in the dataset may potentially be identified in combination with publicly available pictures or videos (from conference attendees or conference official photographer) from other media from the conference the dataset was recorded at. In any case, re-identifying the subjects is strictly against the End User License Agreement under which we share the dataset.

Q. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

We did not request any such information from data participants. Here, the ACM Multimedia '19 General Chair Martha Larson also helped advocate on behalf of the attendees during the survey-design stage. As a result of these discussions, information such as participant gender, ethnicity, or country of origin was not asked.

Q. Any other comments?
None.

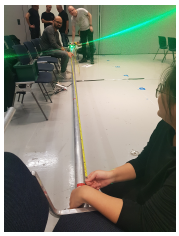
COLLECTION

Q. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The collected data is directly observable, containing video recordings, low-frequency audio recordings and wearable sensing signals (inertial motion unit (IMU) and Bluetooth proximity sensors) of individuals in the interaction scenes. Accompanying data includes self-reported binary categorization of experience level which is available upon request from the authors. The self-reported interests categories are not shared because of privacy concerns.

Video recordings capture the whole interaction floor where the association from multi-modal data to individual is done manually by annotators by referring to frontal (not-shared) and overhead views. The rest of the data was acquired from the wearable sensing badges, which is person-specific (i.e., no participant shared the device). Video and audio data were verified in playback. Wearable sensing data was verified through plots after parsing.

Q. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the s was created. Finally, list when the dataset was first published.



(a) Aligning cameras

(b) Affixing the mounting beam

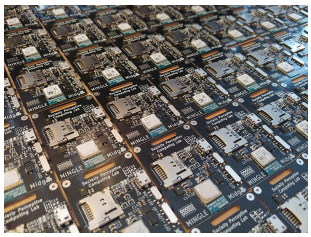
(c) Aligning floor markers

(d) Marking the floor grid

(e) Interaction area

(f) Verifying camera sync.

(g) Assembling Midges



(h) Midges

(i) Verifying crossmodal sync.

Figure 13: Illustrating the process of setting up the data recording.

All data was collected on October 24, 2019, except the self-reported experience level and research interest topics which are either obtained on the same day or not more than one week before the data collection day. This time frame matches the creation time frame of the data association for wearable sensing data. Video data was associated with individual during annotation stage (2020-2021), but all information used for association was obtained on the data collection day.

Q. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

To record videos, we used 14 GoPro Hero 7 Black cameras. The wearable sensor hardware has been documented and open-sourced at https://github.com/TUDelft-SPC-Lab/spcl_midge_hardware. The validation of the sensors was completed through an external contractor engineer. The data collection software was documented and published in [48], which includes validation of the system. These hardwares and mechanisms have been open-sourced along with their respective publication.

The synchronization setup for data collection (intramodal and intermodal) was documented and published in [18], which includes validation of the system.

To lend the reader further insight into the process of setting up the recording of such datasets in-the-wild, we share images of our process in Figure 13.

Q. What was the resource cost of collecting the data?

The resources required to run this first edition of ConfLab include equipment, logistics, and travel costs. Table 5 shows the full breakdown of the costs. The equipment expenses are fixed one-time costs since the same equipment can be used for future iterations of ConfLab. The on-site costs at the conference venue were toward renting a crane for a day to mount the cameras on a scaffold on the ceiling. We have open-sourced the Midge (our custom wearable) schematics so that others don't need to spend on the design and development.

Table 5: Itemized costs associated with recording ConfLab

Item	Cost (USD)
Travel (total for 6 people)	
Flights	1800
Accommodation	1500
Equipment (one time)	
Mounting scaffold	2000
14 $\times$ GoPro Hero 7 Black	4900
Designing the Midge (custom wearable, now made open source)	26000
110 $\times$ Midges (boards, batteries, 4 GB sd cards, cases)	3660
Multimodal synchronization setup	730
Annotations	8000
Computational cost for experiments	500

No additional energy consumption was incurred for collecting the data. However, the ancillary activities (e.g., flights, accommodation) resulted in energy consumption. Flights from the Netherlands to France round-trip for six passengers results in 1020 kg carbon emissions. Accommodation for six members resulted in 22 kWh energy consumption.

Q. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

ConfLab contains both annotated and unannotated segments of multi-modal data. The segment where the articulated pose and speaking status were annotated is selected to maximize crowd density in the scenes. The annotated segment is 16 minutes; the whole set is roughly 1 hour of recordings.

Q. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The Conflab dataset was captured during a special social event called *Meet the Chairs!* at an international conference on signal processing and machine learning. Newcomers and old-timers to the conference freely donated their social behaviour data as part of a *by the community, for the community* data collection effort. Aside from the chance to meet the chairs and create a community dataset, the attendees also received a personalised report of their social behaviour from the wearable sensors (see Appendix C) Conference student volunteers were involved in assisting the set-up of the event. Conference organizers (mentioned in the *Motivation* section) assisted in connecting us with conference venue contacts to mount our technical set-ups in the room. Volunteers and conference organizers were not paid by us. Conference venue contacts were paid by the conference organizers.

Data annotations were completed by crowdsourced workers. The crowdsourced workers were paid \$0.20 for qualification assignment (note that typically requesters do not pay for qualification tasks). Depending on the submitted results, workers earn qualification to access of the actual tasks. The annotation tasks were categorized into low-effort (\$150), medium-effort (\$300), and high-effort (\$450), corresponding to the amount of estimated time each would take. The duration of the tasks was determined by the crowd density and through timing of the pilot studies. The average hourly payment to workers is around \$8.

Q. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

The data collection was approved by the Human Research Ethics Committee (HREC) of our university (Delft University of Technology), which reviews all research involving human subjects. The data collection protocol is also compliant to the conference location's national authorities (France). The review process included addressing privacy concerns to ensure compliance with GDPR and university guidelines, review of our informed consent form, data management plan, and end user license agreement for the dataset and a safety check of our custom wearable devices.

Q. Does the dataset relate to people?

Yes.

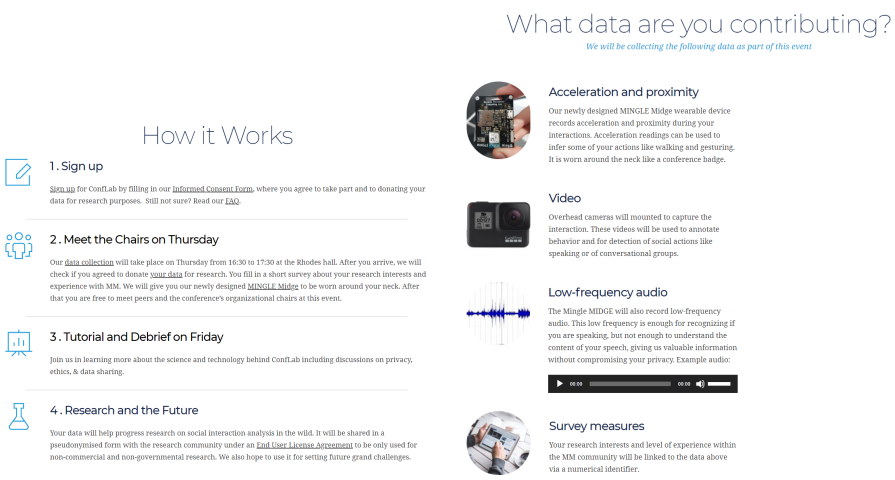


Figure 14: Screenshots of the ConfLab web-page used for participant recruitment and registration.

Q. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

We collected the data from individuals directly.

Q. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

The individuals were notified about the data collection and their participation is voluntary. The data collection was staged at an event called *Meet the Chairs* at ACM MM 2019. The ConfLab web page (`https://conflab.ewi.tudelft.nl/`) served to communicate the aim of the event, what was being recorded, and how participants could sign up. This allowed us to embed the informed consent into this framework so we could keep track of sign ups. See Figure 14 for screenshots. This event website was also shared by the conference organizers and chairs (`https://2019.acmmm.org/conflab-meet-the-chairs/index.html`).

Q. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

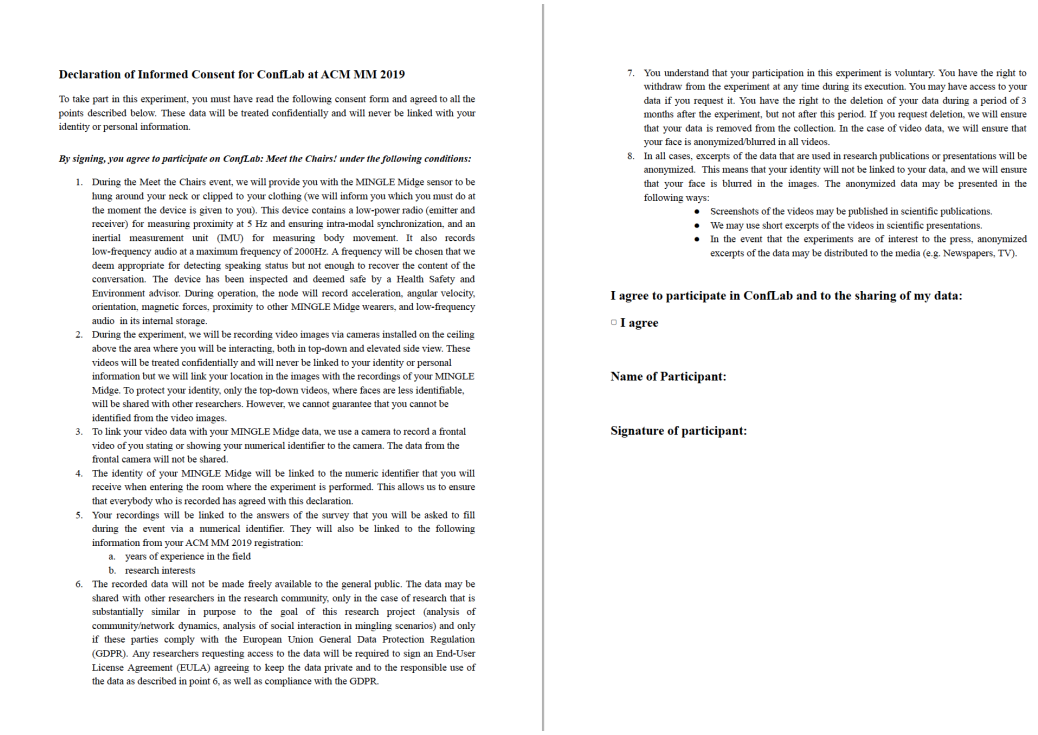
All the individuals who participated in the data collection gave their consent by signing a consent form. A copy of the form is attached below in Figure 15.

Q. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)

Yes, the consenting individuals were informed about the possibility of revoking access to their data within a period of 3 months after the data collection experiment, and not after that. The description is included in the consent form.

Q. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?
No.

Q. Any other comments?
None.



Declaration of Informed Consent for ConfLab at ACM MM 2019

To take part in this experiment, you must have read the following consent form and agreed to all the points described below. These data will be treated confidentially and will never be linked with your identity or personal information.

By signing, you agree to participate on ConfLab: Meet the Chairs! under the following conditions:

1. During the Meet the Chairs event, we will provide you with the MINGLE Midge sensor to be hung around your neck or clipped to your clothing (we will inform you which you must do at the moment the device is given to you). This device contains a low-power radio (emitter and receiver) for measuring proximity at 5 Hz and ensuring intra-modal synchronization, and an inertial measurement unit (IMU) for measuring body movement. It also records low-frequency audio at a maximum frequency of 2000Hz. A frequency will be chosen that we deem appropriate for detecting speaking status but not enough to recover the content of the conversation. The device has been inspected and deemed safe by a Health Safety and Environment advisor. During operation, the node will record acceleration, angular velocity, orientation, magnetic forces, proximity to other MINGLE Midge wearers, and low-frequency audio in its internal storage.
2. During the experiment, we will be recording video images via cameras installed on the ceiling above the area where you will be interacting, both in top-down and elevated side view. These videos will be treated confidentially and will never be linked to your identity or personal information but we will link your location in the images with the recordings of your MINGLE Midge. To protect your identity, only the top-down videos, where faces are less identifiable, will be shared with other researchers. However, we cannot guarantee that you cannot be identified from the video images.
3. To link your video data with your MINGLE Midge data, we use a camera to record a frontal video of you stating or showing your numerical identifier to the camera. The data from the frontal camera will not be shared.
4. The identity of your MINGLE Midge will be linked to the numeric identifier that you will receive when entering the room where the experiment is performed. This allows us to ensure that everybody who is recorded has agreed with this declaration.
5. Your recordings will be linked to the answers of the survey that you will be asked to fill during the event via a numerical identifier. They will also be linked to the following information from your ACM MM 2019 registration:
    a. years of experience in the field
    b. research interests
6. The recorded data will not be made freely available to the general public. The data may be shared with other researchers in the research community, only in the case of research that is substantially similar in purpose to the goal of this research project (analysis of community/network dynamics, analysis of social interaction in mingling scenarios) and only if these parties comply with the European Union General Data Protection Regulation (GDPR). Any researchers requesting access to the data will be required to sign an End-User License Agreement (EULA) agreeing to keep the data private and to the responsible use of the data as described in point 6, as well as compliance with the GDPR.
7. You understand that your participation in this experiment is voluntary. You have the right to withdraw from the experiment at any time during its execution. You may have access to your data if you request it. You have the right to the deletion of your data during a period of 3 months after the experiment, but not after this period. If you request deletion, we will ensure that your data is removed from the collection. In the case of video data, we will ensure that your face is anonymized/blurred in all videos.
8. In all cases, excerpts of the data that are used in research publications or presentations will be anonymized. This means that your identity will not be linked to your data, and we will ensure that your face is blurred in the images. The anonymized data may be presented in the following ways:
    - Screenshots of the videos may be published in scientific publications.
    - We may use short excerpts of the videos in scientific presentations.
    - In the event that the experiments are of interest to the press, anonymized excerpts of the data may be distributed to the media (e.g. Newspapers, TV).

I agree to participate in ConfLab and to the sharing of my data:

□ **I agree**

Name of Participant:

Signature of participant:

Figure 15: Consent form signed by each participant in the data collection.

PREPROCESSING / CLEANING / LABELING

Q. Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

We did not pre-process the signals obtained from the wearable devices or cameras. The only exception is the audio data. Due to a hardware malfunction (this is resolved for the Midges by using different SD cards), the audio needed to be post-processed in order to synchronize it with the other modalities. The synchronization against other modalities was manually checked.

Labeling of the dataset was done as explained in the *Composition* section.

Q. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

The dataset is separated into raw data and the post processed data. For the audio, the original raw data is not suitable for most use cases due to the mentioned synchronization issue. So we share the synchronized version in the raw part of the repository.

Q. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

The processing / fixing of the audio files did not require special software.

The annotation of keypoints and speaking status was done by making use of the Covfee framework: https://josedvq.github.io/covfee/

Q. Any other comments?

None.

USES

Q. Has the dataset been used for any tasks already? If so, please provide a description.

In the main paper, we have benchmarked three baseline 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 for inferring more socially related phenomena. We chose these benchmarking tasks since they have been studied on other in-the-wild behavior datasets.

Q. Is there a repository that links to any or all papers or systems that use the dataset?

None at the time of writing of the paper.

Q. What (other) tasks could the dataset be used for?

Given the richness and the unscripted open-ended nature of the social interactions, ConfLab can be used for many other tasks.

Forecasting, causal relationship discovery Recently, tasks pertaining to the forecasting low-level social cues in conversations have been receiving increased attention from the community [72, 75]. The real-life nature of ConfLab along with the increased data and annotation fidelity can prove a valuable resource for such tasks. Similarly, ConfLab can also be used for efforts towards discovering causal relationships between social behaviors [76].

Data Association. A crucial assumption made in many former multimodal datasets[9, 11, 24] is that the association of video data to the wearable modality can be manually performed. Few works [43, 44] have tried to address this issue but using movement cues alone to associate the modalities is challenging as conversing individuals are mostly stationary. This remains a significant and open question for future large scale deployable multimodal systems. One solution may be to annotate more social actions as a form of top-down supervision. However, detecting pose and actions robustly from overhead cameras remains to be solved.

Conversation floor and F-formation estimation Prior analysis on the MatchNMingle dataset has demonstrated that F-formations can contain multiple simultaneous conversations when the F-formations contain a least 4 people [51]. If this is the case for the ConfLab dataset, this may drastically change how F-formations should be labelled (e.g. returning to being a more subjective task [10]) as more time-precise labelling could enable a more nuanced take on F-formation and conversation floor membership over time.

Multi-class social action estimation More annotations resources were focused on speaker status, F-formation, and keypoint estimation. However, there are a wealth of other social actions in the data that could be interesting to combine into a more complex multi-class social action estimation task. Example social actions include drinking, mobile phone use, hand and head gesture types [9, 77].

Estimation and analysis of socially-related phenomena Beyond the modeling of human behavior which is of interest to the Computer Vision and Machine Learning communities, our benchmarked tasks form the basis for further explorations into downstream prediction of socially-related constructs which is of interest to the Social Science and Social Psychology communities. Such constructs include conversation quality [68, 78], dominance [53], rapport [50], and influence [69].

Investigation of novel crossmodal fusion strategies The baseline tasks in our paper rely only on a late fusion strategy. However, ConfLab's sub-second expected cross modal latency of ~ 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 (for predicting e.g. attraction [79]) which need tolerances as low as 40 ms [18, Sec.3.2]. Prior works coped with lower tolerances by computing summary statistics over input windows [17, 35, 36]. ConfLab enables for the first time, the exploration of Multimodal machine learning approaches for social behaviour analysis

in these highly dynamic in-the-wild settings [65]. Through the provided annotations Conflab also enables research in the topic of usage of mobile phones in small-group social interactions in-the-wild.

Person attribute estimation Estimating individuals that are newcomers/old timers from the dataset may be possible based on their networking strategies.

Q. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Although ConfLab's long-term vision is towards developing technology to assist individuals in navigating social interactions, the data could also affect a community in unintended ways: for instance, cause worsened social satisfaction, a lack of agency, stereotype newcomers and veterans, 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 must 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. Consequently, researchers using the data should be aware that resulting insights may not generalize to the general population.

Q. Are there tasks for which the dataset should not be used? If so, please provide a description.

Beyond the cautionary discussion in the previous question, tasks involving the re-identifying the subjects is strictly against the End User License Agreement under which we share the dataset.

Q. Any other comments?

None.

DISTRIBUTION

Q. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset is available for third parties outside of Delft University of Technology to use for academic research purposes subject signing and approval of our End User License Agreement. The dataset will be hosted by 4TU.ResearchData (see the Maintenance section for description of the 4TU entity).

Q. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be distributed via the 4TU.ResearchData user interface where the data can be downloaded. The dataset has a DOI: https://doi.org/10.4121/c.6034313

Q. When will the dataset be distributed?

The dataset has been available since June 9, 2022.

Q. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset will be distributed under a restricted copyleft license, specified within our End User License Agreement, accessible through the 4TU.ResearchData dataset website. No fees are associated with the license.

Q. Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

Q. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

The terms of our EULA and the European General Data Protection Regulations (GDPR) apply.

Any other comments?
None.

MAINTENANCE

Q. Who is supporting/hosting/maintaining the dataset?

The dataset is hosted by 4TU.ResearchData (https://www.4tu.nl/en/about_4tu/), and supported and maintained by The Socially Perceptive Computing Lab at TUDelft.

Q. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Via email: SPCLabDatasets-insy@tudelft.nl.

Q. Is there an erratum?

No.

Q. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Updates will be done as needed as opposed to periodically. Instances could be deleted, added, or corrected. The updates will be posted on the 4TU.ResearchData dataset website.

Q. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

No limits were communicated to our data participants.

Q. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Only the latest version of the dataset will be maintained. If applicable, we will also host older versions of the data, accessible through the 4TU.ResearchData website.

Q. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We are open to contributions to the dataset. In accordance with our End User License Agreement, contributions should be made available, indicating if there are any restrictions on their contribution. We encourage the potential contributors to contact us to discuss how they wish to be attributed (e.g. citation of a paper or repository related to code/annotations). After finalizing the attribution discussion, we can add the attribution as an update following the same process explained above.

C Sample Participant Report

ACMMM 19 - ConfLab Report

Socially Perceptive Computing Lab - Delft University of Technology

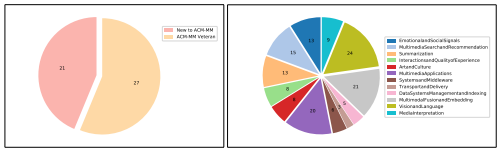
Conflab: Meet the Chairs!

While you were at ACM MM in Nice earlier this year, you had participated in our event called ConfLab: Meet the Chairs!. We want to thank you again for being part of our data collection initiative and contributing to the effort of understanding more about human behaviors and conference experience.

We thought you might be curious about some basic statistics that we have extracted from the collected data. You can find below some general information about all the event participants and some personal information particular to you. Please keep in mind that 1) these are preliminary analyses that we have performed and there could be errors in our estimations, and 2) to protect your privacy, these results are only available to you.

General information about ConfLab participants

When you signed up, we had asked 1) if this was your first time at ACM MM and 2) your research interests (multi-select multiple choice). We had a total of 48 participants. You can see below the statistics over all 48 people.



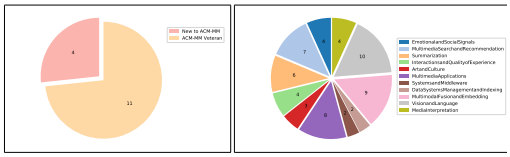
(a) Newcomers and ACMMM Veterans (b) Interests

Figure 1: Statistics of Conflab participants

1

Your networking behaviour - Bluetooth

Here we estimate how many people you have interacted with throughout the event. Our sensors record RSSI values and we set a single threshold for eliminating values corresponding to large physical distance that we do not consider as possible for face-to-face social interactions. We define the criterion of an interaction to be: 1) pairwise RSSI values below -55, and 2) pairwise proximity pings of at least 35 counted within a 1-minute window (sampling rate: 1Hz).



(a) Newcomers and ACMMM Veterans (b) Interests

Figure 2: Statistics of people you interacted with

In Figure 2a, the breakdown of the types of people you have interacted with is shown. In Figure 2b, you will find the interests breakdown of everyone you have interacted with. Figure 3 shows the distribution of the number of participants you interacted with. You will find yourself in the red bin; the x-axis says how many people you have interacted with and the y-axis says how many others had the same numbers as you.

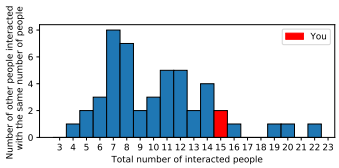


Figure 3: Distribution of the numbers of people participants interacted with

2

Your movement behavior - accelerometer

Here we estimate your motion behavior based on the accelerometer signal. Our sensors record tri-axial accelerometer values and we quantify the amount of motion by calculating the magnitude of the values of all 3 axes. We process the accelerometer data to separate movement and gravitational components of the signals based on a previous approach (Euclidean Norm Minus One [1]). For ease of visualization, we averaged the magnitude of acceleration over 30-second windows. You can see in Figure 4 your personal acceleration magnitude over time, as well as the mean and standard deviation values of acceleration magnitude for all participants over time.

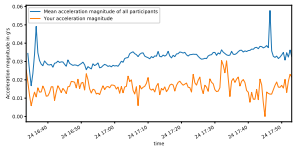
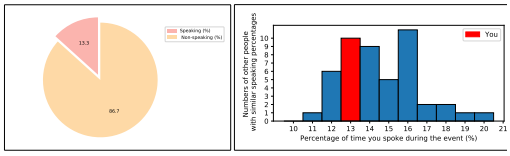


Figure 4: Acceleration magnitudes

Your speech behaviour - low-frequency audio

Here we estimate the amount of time you spoke. We first calculate the envelope of the low-frequency audio signal by taking the absolute value. Then, we apply a moving mean operator to the signal. By manually observing the signals of multiple participants, we selected a threshold to identify the speaking parts of the signal. We then further process the binary stream by filling the gaps between continuous speaking regions and eliminating speech regions that are smaller than a predefined threshold. Figure 5a and 5b show your percentage of speaking during the event and how you compare to the rest of the participants, respectively.

3



(a) Percentages of speech and non-speech (b) Distribution of speaking percentages for all participants

Figure 5: Your speaking behaviour

And that's it from the Socially Perceptive Computing Lab for now!

Note that for us, these analyses are just the starting point for estimating socially relevant behaviours. To do this more robustly and using more complex approaches is one of the reasons why we plan to share the data in next year or so. Maybe you are also curious to develop your own estimation techniques.

Finally, we welcome feedback on what other analyses that you are interested in, technical approaches, how to display your data better, your participatory experience, and any comments or advice that you might have for us. Please feel free to reply to this email or write to one of us directly.

Thanks again for your interest and we hope to see you again in the future!

[1] Bakrania, Kishan, et al. "Intensity thresholds on raw acceleration data: Euclidean norm minus one (ENMO) and mean amplitude deviation (MAD) approaches." PloS one 11.10 (2016): e0164045.

4

Figure 16: Sample post-hoc report sent to each participant of ConfLab. The report contains insights into the participant's networking behavior from the collected wearable-sensors data. This insight served as an additional incentive to participate in ConfLab, beyond interacting with the Chairs and contributing to a community-driven data endeavor (see main paper Section 3).

D Data Capture Setup Details

The Midge We improved upon the Rhythm Badge in three ways towards enabling more fine-grained and flexible data capture: (i) enabling full audio recording with a frequency up to 48 KHz, with an on-board switch to allow physical selection between high and low frequency capture directly at acquisition; (ii) adding a 9-axis Inertial Measurement Unit (IMU) with an on-board Digital Motion Processor (DMP) to record orientation; and (iii) an on-board SD card to directly store raw data, avoiding issues related to packet loss during wireless data transfer required by the Rhythm Badge. IMUs combine three tri-axial sensors: an accelerometer, a gyroscope, and a magnetometer. These measure acceleration, orientation, and angular rates respectively. These sensor measurements are combined on-chip by a Digital Motion Processor. Rough proximity estimation is performed by measuring the Received Signal Strength Indicator (RSSI) for Bluetooth packets broadcast every second (1 Hz) by every Midge. During the event, IMUs were set to record at 50 Hz. We recorded audio at 1250 Hz to mitigate extraction of verbal content while still ensuring robustness to cocktail-party noise.

Wireless Synchronization at Acquisition The central idea for our syncrhonization approach involves using a common Network Time Protocol (NTP) signal as reference for the camera and wearables sub-networks. The set-up achieved a cross-modal latency of 13 ms at worst, which is well below the 40 ms latency tolerance suitable for behavior research in our setting [18, Sec. 3.3]. Additionally, our synchronization approach allowed for dynamic addition of sensors to the network while still obtaining synchronized data streams. This is crucial in extreme in-the-wild events where some participants might arrive late.

Sensor Calibration For computing the camera extrinsics, we marked a grid of 1 m $\times$ 1 m squares in tape across the interaction area floor. We ensured line alignment and right angles using a laser level tool (STANLEY Cross90). For computing the camera intrinsics, we used the OpenCV asymmetric circles grid pattern [80]. The calibration was performed using the Idiap multi camera calibration suite [81]. All wearable sensors include one TDK InvenSense ICM-20948 IMU [82] unit that provides run time calibration. To establish a correspondence with the camera frame of reference, the sensors were lined up against a common reference-line visible in the cameras to acquire an alignment so that the camera data can offer drift and bias correction for the wearable sensors.

E Implementation Details

E.1 Person and Keypoint Detection Models

Data Cleaning A few frames contained some incorrectly labeled keypoints, a product of annotation errors like mis-assignment of participant IDs. We removed these using a threshold on the proximity to other keypoints of the same person. Further, in some cases, a person might be partially outside a camera's field of view. For the person detection task, we compute the bounding box from the keypoint ground-truth annotations. If more than half the body (50% keypoints) is missing in the frame so that e.g. only their legs are visible (see top of Figure 7a), we don't consider the person for that frame in the person detection experiments. Note that due to the significant overlap between the camera views, the person would be considered for the corresponding frame in the next camera. If they move back into the original view, we again take them into consideration for the original camera for the corresponding frame. Moreover, if there are more than 10% missing keypoints across all people in an image, we also discard that image from the experiment. This preprocessing resulted in a training set with 112k frames (1809k person instances) and a test set with 7k frames (158k person instances).

Training We resized the images to 960×540, and augmented the data by randomizing brightness and horizontal flips. The learning rate was set to 0.02 and batch size to 4. We trained the models for 50 k iterations, using the COCO-pretrained weights for initialization. All hyper-parameters were chosen based on the performance on a separate hold-out camera chosen as validation set. During training, any missing ground-truth keypoints (resulting from the person being partially outside the camera's view for instance) are ignored during back-propagation.

E.2 F-formation Detection

Data Cleaning Because keypoint annotations of the subjects are based on camera view and that the F-formation clustering methods cannot group subjects that do not exist under one camera view (e.g., when there are more identities than in associated ground truths), we processed the ground truth also based on camera number. This filtering pre-processing was decided based on the best camera view of the F-formations.

Feature Extraction The required features of GCFF and GTCG include location and orientation of the subjects. We used the X and Y position of subjects' head (as it is the most visible from the top-down view) for location, and extracted orientations for head, shoulders and hips. The orientations are calculated based on corresponding vectors determined by head and nose keypoints, left and right shoulder keypoints, and left and right hip keypoints, respectively.

Training We used pre-trained parameters for field of view (FoV) and frustum aperture (GTCG) and minimum description length (GCFF), provided in these models trained on the Cocktail Party. FOV and aperture are related to human eye gaze and head anatomical constraints reported by [83], and hence not dataset specific. The minimum description length is an initialized prior dictated by the same form of the Akaike Information Criterion, and becomes part of the optimization formulation. We tuned parameters such as frustum length (GTCG) and stride (GCFF) to account for average interpersonal distance in ConfLab based on Camera 6, as they vary across different datasets.

F Additional Results

F.1 Person and Keypoints Detection

Predictions from pretrained SOTA models Figure 17 shows predictions from SOTA human keypoint estimation models, namely, RSN [19], MSPN[84], HigherHRNet [85], and HourglassAENet [86], for the testing images of the Conflab dataset. Note that RSN and MSPN are top-down networks, i.e., they require person bounding boxes to predict the keypoints in each bounding box. We use COCO pretrained faster-RCNN network for bounding box estimation. HigherHRNet and HourglassAENet are bottom-up models, i.e., they directly predict keypoints from the full image. We use publicly available COCO pretrained checkpoints for prediction. The results show that the *state-of-the-arts 2D body keypoint detection models fail to capture the body keypoints in the Conflab dataset*. We infer that training on the dataset (e.g., COCO) that contains mostly side-view images does not work well in top-view images, for which Conflab dataset is important to the community.

Qualitative Results from ResNet-50 Finetuning Figure 18 illustrates more qualitative results from our finetuning experiments. We find that finetuning on our non-invasive top-down camera perspective significantly improves the keypoint estimation performance.

Ablations Tables 6 and 7 include the results of our experiments investigating the effect of varying the training data size on keypoint detection performance (see main paper Section 6.1). In Table 8, we show keypoint detection scores for experiments with different number of keypoints. We first focus on the five upper body keypoints: {head, nose, neck, rightShoulder, leftShoulder}. We then additionally considered the torso region keypoints for a total of nine: {rightElbow, rightWrist, leftElbow, leftWrist}. Finally, we add the hip keypoints {rightHip, leftHip} to the set. The experiments in the main paper are performed with all 17 keypoints. The results show that performance drops slightly when adding the arms keypoints ($5 \rightarrow 9$, AP^{OKS}_{50} and AP^{OKS}), and that the relative gain when adding the hip keypoints ($9 \rightarrow 11$) is lower than when adding the lower body keypoints ($11 \rightarrow 17$, especially AP^{OKS}_{75}). We believe this is largely due to the lower body being more static relative to the arms that move a lot to execute gestures during conversations.

F.2 Speaking Status Detection

Experiments with different sensor modalities Table 9 displays the results from experiments using specific modalities from our IMUs for the task of speaking status detection. We used the best performing classifier (Minirocket [64]) among the ones tested in Table 3. The experiment setup is the

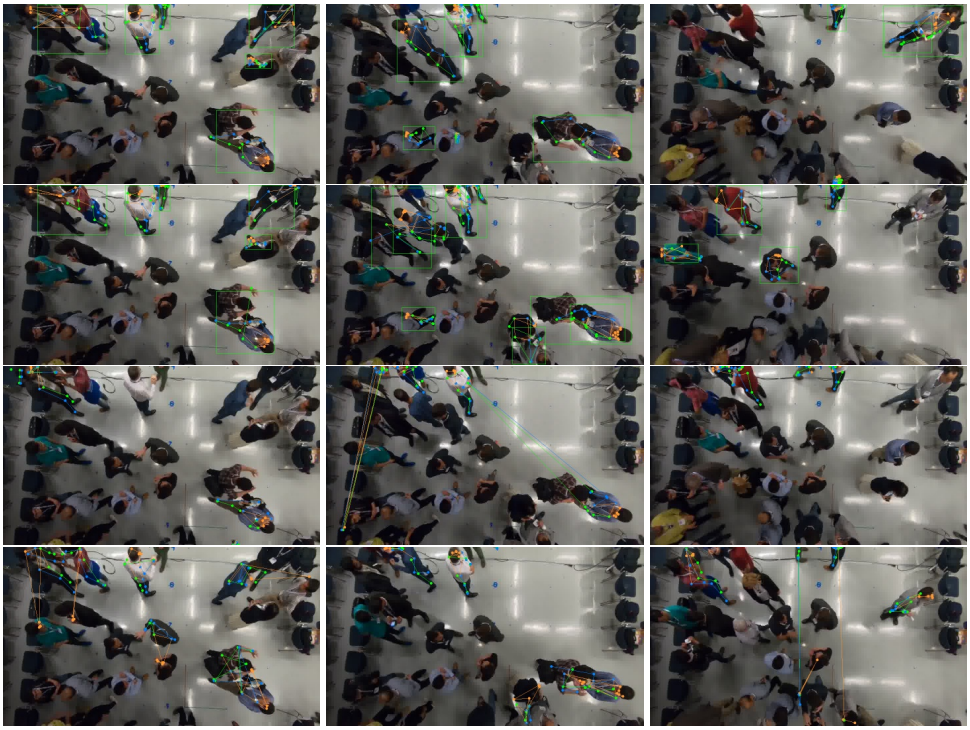


Figure 17: Results from Pretrained keypoint detection models. From top to bottom - predictions from RSN [19], MSPN[84], HigherHRNet [85], and HourglassAENet [86]. Results show that *SOTA 2D body keypoint detection models fail to capture the body keypoints in the ConfLab dataset*.

Figure 18: Results from (top) COCO pretrained Mask-RCNN model, (bottom) our ConfLab finetuned Mask-RCNN model.

Table 6: Effect of varying % frames from each camera at training on keypoint estimation.

% of training samples	AP^{OKS}_{50}
1.6%	29.0
3.2%	35.9
8%	39.0
16%	44.5
100%	45.3

Table 7: Effect of adding all frames from individual cameras to the training set on keypoint estimation.

Train Camera	#(training samples)	AP^{OKS}_{50}
cam 2	34k	8.6
cam 2 + cam 4	69k	31.1
cam 2 + cam 4 + cam 8	112k	45.3

Table 8: Keypoint estimation ablation with keypoints from different body sections: head and shoulders (5), + torso (9), + hips (11), + knees and feet (full 17).

#Keypoints	AP^{OKS}_{50}	AP^{OKS}	AP^{OKS}_{75}
5	26.6	7.1	1.4
9	26.5	6.9	2.0
11	35.8	9.5	2.2
17	45.3	13.5	3.3

Table 9: ROC AUC and accuracy for different sensor modalities from out 9-dof IMU in speaking status detection using the Minirocket classifier [64]. The number of channels in the corresponding modality is indicated in parentheses.

Input Modality	AUC	Accuracy
Acceleration (3)	0.813	0.768
Gyroscope (3)	0.765	0.716
Magnetometer (3)	0.610	0.656
Rotation vector (4)	0.726	0.696
All (13)	0.774	0.739

same as detailed in Section 6.2, and the model is not changed between runs, except for the fact that different modalities may have a different number of input channels.

G Reproducibility Checklist

G.1 Person and Keypoints Detection

- Source code link: https://github.com/TUDelft-SPC-Lab/conflab
- Data used for training: 112k frames (1809k person instances).
- Pre-processing: See Section 4, Appendix E.1.
- How samples were allocated for train/val/test: cameras 2, 4, and 8 are selected for training. For hyperparameter tuning, camera 8 are held out for validation.
- Hyperpatameter consideration: We considered learning rates $(0.001/0.005/0.05/0.01)$, number of epochs $(10/20/50/100)$, detection backbone (R50-FPN/R50-C4). Also see Appendix E.1
- Number of evaluation runs: 5
- How experiments were ran: See Section 6.1.
- Evaluation metrics: Average precision at different thresholds.
- Results: See Section 6.1 and Appendix F.1.
- Computing infrastructure used: All baseline experiments were ran on Nvidia V100 GPU (16GB) with IBM POWER9 Processor.

G.2 Speaking Status Detection

- Source code link: https://github.com/TUDelft-SPC-Lab/conflab
- Data used for training: 42884 windows (3 seconds), extracted from 48 participants' wearable data and speaking status annotations
- Pre-processing: Data was windowed into 3-second segments (see Section 6.2). The source code includes this pre-processing step.
- How samples were allocated for train/val/test: 10-fold cross-validation at the subject level (48 subjects) to test generalization to unseen data subjects. The splits can be reproduced exactly using the source code.
- Hyperparameter considerations: For acceleration-based methods, we used default network hyper-parameters and architectures from their tsai implementation [87]. For the MS-G3D baseline [61], we used default hyperparameters from the authors' implementation. For both, we determined the early stoppage point using a small subset (10%) of the training set.
- Number of evaluation runs: 1 run of 10-fold cross-validation
- How experiments were ran: For each fold, the early stoppage point was first determined using 10% of the training data as validation set and AUC as performance metric. The model at this stoppage point was then applied to the test set for evaluation.

- Evaluation metrics: Area under the ROC curve (AUC)
- Results: See Section 6.2
- Computing infrastructure used: Experiments were ran on a personal computer with GPU acceleration (NVidia RTX3080).

G.3 F-formation Detection

- Source code link: https://github.com/TUDelft-SPC-Lab/conflab
- Data used for training: Camera 6
- Pre-processing: See Section E.2 for data cleaning and feature extraction.
- How samples were allocated for train/val/test: samples from Camera 6 were used to select the best model parameters. The rest are for test (evaluation). However, we note that Table 4 shows averaged performance on all cameras to provide a holistic view of the F-formation detection performance on ConfLab.
- (Hyper)parameter considerations: Both baseline methods are not deep-learning based and model parameters are interpretable. For GTCG, the parameters are frustum length (275), frustum aperture (160), frustum samples (2000), and sigma for affinity matrix (0.6). For GCFF, the parameters are minimum description length (30000) and stride (70).
- Number of evaluation runs: 1
- How experiments were ran: A total of eight experiments were run for choosing the best parameters, and three for evaluation (for camera 2, 4, and 8). The parameters were chosen based on grid-search. For optimizing frustum length in GTCG, we searched over $[170, 195, 220, 245, 275]$ with 275 being averaged interpersonal distance based on Camera 6. For optimizing stride D in GCFF, we searched over $[30, 50, 70]$.
- Evaluation metrics: F1
- Results: See Section 6.3
- Computing infrastructure used: The experiments were run on Linux-based cluster instances on CPU with Matlab 2018a.