The Unobtrusive Group Interaction (UGI) Corpus

Indrani Bhattacharya
Rensselaer Polytechnic Institute
Troy, New York
indrani.electrical@gmail.com

Michael Foley
Northeastern University
Boston, Massachusetts
foley.mic@husky.neu.edu

Christine Ku
Rensselaer Polytechnic Institute
Troy, New York
kuc3@rpi.edu

Ni Zhang
Rensselaer Polytechnic Institute
Troy, New York
zhangn5@rpi.edu

Tongtao Zhang
Rensselaer Polytechnic Institute
Troy, New York
zhangt13@rpi.edu

Cameron Mine
Rensselaer Polytechnic Institute
Troy, New York
minec@rpi.edu

Manling Li
Rensselaer Polytechnic Institute
Troy, New York
limanlingcs@gmail.com

Heng Ji
Rensselaer Polytechnic Institute
Troy, New York
jih@rpi.edu

Christoph Riedl
Northeastern University
Boston, Massachusetts
c.riedl@neu.edu

Brooke Foucault Welles
Northeastern University
Boston, Massachusetts
b.welles@neu.edu

Richard J. Radke
Rensselaer Polytechnic Institute
Troy, New York
rjradke@ecse.rpi.edu

ABSTRACT

Studying group dynamics requires fine-grained spatial and temporal understanding of human behavior. Social psychologists studying human interaction patterns in face-to-face group meetings often find themselves struggling with huge volumes of data that require many hours of tedious manual coding. There are only a few publicly available multi-modal datasets of face-to-face group meetings that enable the development of automated methods to study verbal and non-verbal human behavior. In this paper, we present a new, publicly available multi-modal dataset for group dynamics study that differs from previous datasets in its use of ceiling-mounted, unobtrusive depth sensors. These can be used for fine-grained analysis of head and body pose and gestures, without any concerns about participants' privacy or inhibited behavior. The dataset is complemented by synchronized and time-stamped meeting transcripts that allow analysis of spoken content. The dataset comprises 22 group meetings in which participants perform a standard collaborative group task designed to measure leadership and productivity. Participants' post-task questionnaires, including demographic information, are also provided as part of the dataset. We show the utility of the dataset in analyzing perceived leadership, contribution, and performance, by presenting results of multi-modal analysis using our sensor-fusion algorithms designed to automatically understand audio-visual interactions.

MMSys '19, June 18–21, 2019, Amherst, MA, USA

ACM ISBN 978-1-4503-6297-9/19/06…$15.00
https://doi.org/10.1145/3304109.3325816

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing systems and tools**; • **Applied computing** → *Psychology*;

KEYWORDS

Multimodal dataset; multimodal sensing; time-of-flight sensing; face-to-face group interactions; computational social psychology; multimodal interaction.

ACM Reference Format:
Indrani Bhattacharya, Michael Foley, Christine Ku, Ni Zhang, Tongtao Zhang, Cameron Mine, Manling Li, Heng Ji, Christoph Riedl, Brooke Foucault Welles, and Richard J. Radke. 2019. The Unobtrusive Group Interaction (UGI) Corpus. In *10th ACM Multimedia Systems Conference (MMSys '19), June 18–21, 2019, Amherst, MA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3304109.3325816

1 INTRODUCTION

Groups are an intriguing social phenomena that form the core of any organization's functioning. Kurt Lewin [26, 27] coined the term group dynamics and defined it as the the verbal and non-verbal behavior and psychological processes that occur within or between groups. Group dynamics can change even at a millisecond time frame, making them notoriously difficult to identify and measure [11, 23, 24]. Researchers studying group dynamics need to account for the temporal and spatial resolution of events occurring in a group at a very fine granularity; this translates to huge amounts of data. For example, the authors in [25] reported that an analysis of the verbal communication in a one-hour team meeting required approximately 7 hours of intense human coding effort. Manually coding nonverbal behavior like location, gaze, gestures, head and body movements, posture, speaker segmentation, interruptions and so on, is even more cumbersome and time-consuming.

The availability of multi-modal sensors and advances in computer vision, machine learning, and natural language processing make it possible to automatically analyze several behavior patterns that are important for the automatic analysis of social interaction. Gatica-Perez [12] reviewed around a hundred papers dealing with small social interactions with a focus on non-verbal behavior, computational models, social constructs, and face-to-face interactions. The fine-grained analyses of group interaction patterns using these automated methods help in understanding social constructs such as agreement/disagreement [6], cohesion [15], dominance [16], leadership [3, 18, 38] and emotion [30] in group interactions.

Although many automated methods to analyze human behavior exist, there are only a few publicly available multi-modal datasets that can simultaneously help both communities of computer scientists and social scientists [25]. Several researchers from the domains of social psychology, social signal processing, multimodal interaction, and affective computing have expressed the need for publicly available multi-modal datasets that enable research in this interdisciplinary field [32]. We are also motivated by studying group interactions where the participants are comfortable and uninhibited, necessitating unobtrusive sensing techniques. Frontal video cameras or wearable sensors can defeat the purpose of studying natural human behavior.

In this paper, we present a new small-group interaction dataset called the UGI Corpus (for Unobtrusive Group Interaction) whose primary novelty is the use of ceiling-mounted depth sensors in an unobtrusive and identity-preserving manner, which we believe to be the first of its kind. Despite their overhead angle and lack of color information, ceiling-mounted depth sensors allow for the automatic identification of head pose and visual focus of attention (VFOA), upper body and head movement analysis, and body posture and arm pose estimation. We also collected synchronized audio information and automated, anonymized, time-stamped transcripts of the group meetings. Our dataset additionally contains demographic information and the results of a post-task participant questionnaire that allows the correlation of derived multimodal metrics with emergent leadership and contribution. In the rest of the paper, we discuss the existing datasets for studying face-to-face group interactions, our sensing infrastructure and data collection methods, our automated algorithms for audio-visual understanding of human behavior, and our preliminary analyses of the dataset. The dataset is summarized and is publicly available for download at https://sites.google.com/view/ugirpi/.

2 RELATED WORK

Among the early research efforts to study face-to-face group interactions, the most prominent is work from the Human Dynamics Research Group at MIT, who explored the use of wearable sensors such as the "SocioMeter" [10] and the "sociometric badge" [22, 34]. The sociometric badge could be used to understand (1) common daily activities like sitting, standing, walking, and running in real time using a 3-axis accelerometer, (2) social signals like interest, excitement, and interjections from extracted speech features, (3) location to an accuracy of 1.5m by measuring received signal strength and using triangulation algorithms, and (4) whether people wearing the badges are facing one another within a 30 cone and 1m distance by using an IR sensor.

The ICSI [17] and ISL [8] meeting corpora contain audio data from several natural and scripted meetings collected with the aim of facilitating research in automatic speech recognition, noise robustness, dialogue modeling, transcription, prosody and speaking styles. The AMI corpus [20] was collected in the IDIAP smart room for studying addressing behavior in small, face-to-face conversations. It contains multi-modal sensor data and hand-annotated meeting dialogues, gaze directions, addressees, and adjacency pairs. The ATR database [9] collected meeting data using a table-top sensor device consisting of a small 360-degree camera surrounded by an array of high-quality directional microphones. The Mission Survival (MS-2) corpus [29] contains audio and video recordings of group discussions on a hypothetical plane crash scenario. This corpus was developed to study personality traits and social behavior using audio-visual cues. The ELEA corpus [36] was formed with the goal of analyzing emergent leadership in newly formed groups. Close-talking mono-directional microphones, Microsoft Kinects, and GoPro cameras were used to build the corpus [33] for analyzing conversational behavior in group interviews.

A recent project by Müller et al. for understanding rapport in conversations [31] is based on a multi-modal dataset in which each participant is recorded by two external cameras and an omni-directional microphone. Braley and Murray presented the Group Affect and Perfomance (GAP) corpus [7], containing thirteen small-group interactions in which the participants perform the Winter Survival Task. The Winter Survival task [19, 21] is a group decision making task where the participants discuss critical survival items in a plane crash scenario in the woods during a severe winter. In the GAP corpus, the meetings were recorded with a portable audio recorder placed in the center of the group members, with a webcam in front of each participant to record the frontal upper body view. The publicly available dataset from this corpus contains audio recordings, meeting transcripts, and post-task questionnaire answers that included demographic details and perceptions on cohesion, efficiency, time management, and leadership.

The available datasets for studying group behavior depend heavily on the use of special wearable sensors [22, 34, 35], one or more cameras [4, 18], or front-facing Kinects [33]. The presence of visible cameras can alter participants' natural behavior [39], and having unusual sensors directly in front of ones' face or in the line of sight may inhibit natural group interactions. Also, datasets with frontal cameras that reveal the identity of the participants are often more difficult to make publicly available to the research community. We posit that a room in which the participants are as unaware of being sensed as possible, and where the data collection approach does not intrude into the identity and privacy of the participants, is beneficial for studying natural group dynamics.

As opposed to existing datasets that capture frontal face-to-face interactions of small groups, we present a multi-modal dataset of task-based small group interactions, using unobtrusive sensing techniques. We recorded 22 group meetings of 86 participants, with group size ranging from 3–5, using ceiling-mounted Kinect depth sensors and individual lapel microphones. The depth sensors are out of the lines of sight of participants and are privacy-preserving. Participants completed the Lunar Survival Task [14] (described in Section 3.1), and filled out a post-task questionnaire on their demographics and perceptions of leadership, contribution, and the

nature of discussion. We publish the synchronized overhead depth videos, the anonymized, time-stamped meeting transcripts, and the post-task questionnaire answers. We do not publish the speech signals, as these may contain spoken names (which are anonymized in the transcript), and can also reveal the identity of the participants. We also summarize our baseline algorithms for automated analysis on this multi-modal data and our preliminary results.

3 THE UGI CORPUS: SENSING INFRASTRUCTURE

We conduct our experiments in a specially modified 11^0 28^0 conference room. Figure 1 shows the meeting room, the two kinds of sensors we use, and the group meeting layout. The visual aspects of the meeting are captured by the depth sensors of 2 ceiling-mounted Microsoft Kinects, placed above the edges of the table. The audio information is captured by lapel microphones on each participant.

We record only the depth information from the Kinects, and do not save any RGB information. The depth sensor of the Kinect operates on the principle of time-of-flight (ToF) of light. Figure 2 shows the depth map from each of the 2 Kinects with associated person tags, capturing the participants on each side of the table. It is clear from Figure 2 that the overhead depth information alone does not reveal the identity of the participants and hence their anonymity is preserved. Since the ToF sensors are embedded in the ceiling, they are outside participants' sight lines and there is no sense of being "watched". In addition to preserving the identity of the participants and being unobtrusive, the ToF sensors have two advantages compared to cameras. First, they return distance maps instead of images, enabling the direct creation of 3D point clouds of the environment, and second, they are more robust to variations in the ambient lighting in the environment and the color/reflectiveness of the participants' clothing.

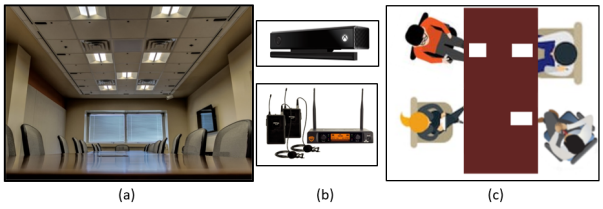

Figure 1: (a) The instrumented meeting room, which contains (b) two ceiling-mounted Microsoft Kinects and per-participant lapel microphones. (c) The layout of a typical meeting.

3.1 Task-based interaction

3.1.1 The Lunar Survival Task. We recorded 86 individuals across 22 groups who completed the Lunar Survival Task [14] in the conference room using the sensing architecture described in Section 3. The Lunar Survival Task is a widely-used group discussion task that assesses the effects of deliberation processes on decision-making quality. In small groups of 3–5, participants discuss a hypothetical survival scenario on the moon and rank the value of 15 supplies that may aid in their survival.

The first stage of the task requires the participants to individually rank the 15 items in order of their importance, without communicating with the other members of the group. This stage takes 10 minutes. In the second stage, the participants work as a group to rank the items. The group has to employ the method of group consensus in reaching its decision within a maximum of 15 minutes. Since each item ranking must be agreed upon by all the group members, this task requires collaboration in order to reach consensus. Further, some members can guide the discussion, influence the rankings more than the other members, and act as emergent leaders, although no group leader is designated. All the discussions were in English.

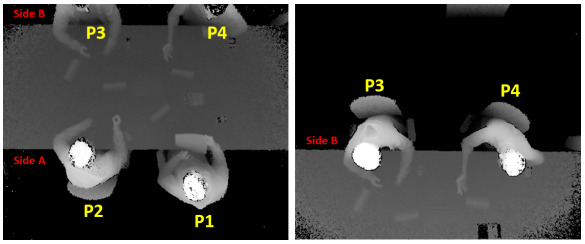

Figure 2: The depth map from each overhead Kinect with associated person tags. One Kinect captures individuals seated on side A of the table completely and individuals on the opposite side B of the table partially. The other Kinect completely captures individuals on side B of the table only. Raw undistorted depth measurements are pre-processed for maximum dynamic range.

Before the beginning of each task, the participants were made aware of the existing sensors and that they would be recorded using the overhead Kinects, the microphones, and two reference video cameras. If they felt comfortable with participating in the study, they were asked to sign a consent form that detailed the task and the sensing infrastructure prior to the beginning of the experiment.

3.1.2 The Post-Task estionnaire. After the group discussion, each participant is asked to complete a post-task questionnaire. In addition to questions relating to the age, race, and gender of the participants and whether they had ever completed the task before, the post-task questionnaire also asked the participants to rate on a 5-point scale (not at all, a little, somewhat, a lot, a great deal) the following questions:

> How well did you know each of your group members before today?
> Before today, how familiar were you personally with the topic of survival in space?
> To what extent did the following group members contribute to the discussion? [28]
> To what extent did the following group members act as a group leader?
> For each of the following pairs of words, please indicate the point on the scale that best represents your feelings about the group conversation: engaging–boring, warm–cold, comfortable–awkward, interesting–dull, friendly–detached. [2]

3.1.3 Demographics. The participants were all undergraduate or graduate students. From the post-task questionnaire data, 34 participants self-identified themselves as 'White', 40 participants self-identified themselves as 'Asian', 10 participants self identified as 'Hispanic' and 2 participants self-identified as 'White/Hispanic'. The age of the participants varied from 18 years to 29 years, with a mean age of 21 years, median age of 20 years and mode 18 years. Based on self reports, 51 participants were men and 35 participants were women.

4 DATA PREPROCESSING

4.1 Synchronizing the different modalities

The overall recorded multi-modal data included the depth map from the 2 overhead Kinect sensors (at 15 fps), and audio information collected from individual lapel microphones on each participant (at 48kHz). We also collected reference video using two video cameras at the far ends of the room. The video camera data is not used for any algorithm development and is purely used for illustrations and ground truth determination. The video camera data is not part of the UGI corpus or being released with this dataset.

In order to synchronize the different modalities, each meeting started with a clap from a non-participant, visible to one Kinect and audible in all microphones. The two Kinects are synchronized by a single host PC. All sensors could thus be synchronized using the clap as the audio-visual cue for the start of the meeting. The synchronization process involved the following steps:

Synchronizing audio and reference video:

Both the audio and video files were opened in Audacity and the clap (which is our reference sound) was precisely located in both modalities to a millisecond resolution of accuracy.
The audio and video tracks were independently shifted backwards to have the reference clap as the starting position of each track, by deleting information before the clap.
Once shifted, the alignment of audio and video tracks was checked by playing both modalities together.

Synchronizing audio and Kinect depth information:

The precise location of the clap was determined by observing the overhead depth maps from the Kinects.
Each Kinect frame has an associated time-stamp. The time-stamp of the clap frame became the zero position of the shifted Kinect video.
The Kinect frames are recorded at 15fps, while the audio recording is at 48kHz. In order to have completely synchronized audio-Kinect data, we needed to down-sample the audio frames to the Kinect frame-rate. This was done by noting the time-stamp differences between subsequent frames of the Kinect starting from the clap, and calculating the corresponding audio frame number using this time difference.

4.2 Preprocessing the distance maps

The Kinect depth sensing range is approximately 0–6m. In order to get the best dynamic range for algorithm development, we performed the following preprocessing steps on the overhead depth map:

The distance measurements were subtracted from the room height (2.8m), so that distance measurements are made from the floor upwards as opposed to the raw measurements from the ceiling downwards.
We noted from the depth maps of our dataset that seated individuals are not further than 1.3m from the ground. Also, distance information below 0.6m is not very useful, since the lower torsos of the participants are covered by the table (0.9m high). In order to have the maximum dynamic range, we clipped all measurements below 0.6m from the floor level to 0.6m and all measurements above 1.3m to 1.3m.
We performed linear stretching of the distance measurements in the range of 0.6–1.3m to the grayscale range [0, 255], and formed videos at 15fps with these pre-processed distance maps. Snapshots of the Kinect distance maps after these pre-processing steps are shown in Figure 2.

4.3 Non-verbal speech segmentation

We post-processed the aligned, recorded audio signals using noise reduction in Audacity [1]. We performed speaker segmentation and silence detection on these synchronized, noise-reduced audio signals using techniques described in [13]. Essentially, for each lapel microphone recording, speech segments were detected by applying a dynamically estimated thresholding criterion on the extracted signal energy and the spectral centroid. As described in Section 4.1, accurate timestamps also allowed us to downsample the speaker identification information (collected at 48kHz) to the Kinect ToF frame rate of 15 fps.

4.4 Transcription

The recorded audio was transcribed to text using IBM Watson's Speech-to-Text API [37], which uses Long Short-Term Memory (LSTM) and Residual (ResNet) neural networks. In order to boost the automatic transcription performance, we performed two preprocessing steps on the audio signal. First, the synchronized audio track corresponding to each individual was independently segmented to distinguish between speech and silence sections, as described in Section 4.3, and then the silence sections were zeroed out. Next, we selected short segments of 1–2 minutes of each individual audio track, such that the corresponding individual has a speaking section during this short segment. The IBM speech-to-text transcription software was then run several times on this short segment, with different values of amplitude threshold, to determine the threshold value that gives a resulting transcript that picks up the maximum spoken content of the concerned individual and also has accurate timestamps corresponding to a maximum number of transcribed lines.

After the threshold has been selected for each track individually, all the tracks are sent through the IBM speech-to-text transcription module, which returns individual transcripts for each person. Finally, each of the individual transcripts was passed through another script that sorts the time-stamps and accurately merges the transcripts. Each person is anonymized as Person 1, Person 2, etc., and named mentions of individuals (if any) are also anonymized. Once the threshold values are decided for each file, the entire process of transcription, including individual transcription and merging the multi-person transcription, takes approximately 3 minutes for one

minute of 4-person discussion, while running on a Windows 10 computer, with 32GB RAM and Intel(R) Xeon(R) E5-2620 v3@2.40GHz processor. Therefore, for a 4-person 15-minute discussion, the automatic transcription process requires around 45 minutes. After the automatic transcription, each transcription file is manually touched up to ensure correctness of the transcription process.

To evaluate the transcription performance, we adopted the widely used Word Error Rate (WER). We computed the WER as 26.61% on average for five randomly selected meetings, since oral representations are informal and rich in discourse markers. Since transcription is the foundation for verbal-speech-related secondary tasks, we also analyzed the effect of transcription performance on these tasks. Considering that words have widely varying weights for secondary tasks (e.g., words containing knowledge are more important than others and less errors are allowed), we considered the difference in information extraction performance between the automatic transcripts and touched-up transcripts. In this dataset, the knowledge is mainly about the lunar survival task items. Thus, we extracted the item and compared extraction results. The average precision difference is 2.37%, showing that the touched up transcriptions have the ability to provide support for secondary tasks.

5 AUDIO-VISUAL UNDERSTANDING FROM THE UGI CORPUS

Our primary purpose in this paper is to propose and disseminate the UGI Corpus. In the following sections, we discuss the utility of the dataset in studying emergent leadership, contribution and performance in task-based group interactions, by briefly describing our automated algorithms for multimodal understanding and preliminary analyses.

Visual Understanding: The overhead depth information from the two Kinects allows us to accurately compute 3D point clouds enabling tracking of participants, understanding their coarse body and head poses, and classifying visual focus of attention (VFOA) target locations. In order to estimate the VFOA of each participant at each instant of time, we developed a multi-sensor fusion algorithm that leverages the depth information to estimate the head pose and the synchronized speaker identification information to derive a contextual understanding of the meeting at that point of time [5]. The VFOA classifier gives the VFOA target location for each participant as one of the other participants, the paper in front of the participant, or "unfocused". We achieved an overall VFOA classification accuracy of 48%, which is comparable to accuracies using front-facing cameras (42%) in similar group meeting settings [18]. A short video clip illustrating the VFOA estimation on a meeting segment is at https://youtu.be/s1yaZk3hKFY.

Non-verbal Speech Understanding: We use the synchronized and segmented speech signals to understand individual non-verbal metrics like speaking length, successful and unsuccessful interruptions, speaking turns, back-channels, and group level metrics such as group silence, and overlapping speaking lengths [5].

Verbal Speech Understanding: We use Natural Language Processing (NLP) algorithms on the meeting transcripts to detect and extract individual opinions as the discussion proceeds, and also to understand what influence each participant has on the other members in reaching the consensus. The algorithm picks up explicit item and ranking mentions, and agreements/disagreements, and constructs a bipartite graph that captures the current state of the meeting [5, 40]. A short video clip illustrating this graphical summarization can be found at https://youtu.be/asLSE1pxTFk.

In order to study the ability of these automatically extracted metrics to study human perceptions of leadership and contribution, we extracted 20 audio-visual metrics at the individual level, including the amount and ratio of visual attention received and given by each participant, the speaking length, turns and interruption patterns, and the role of each participant in proposing items and ranking, summarizing discussions, and introducing new relevant information [5]. We studied the correlation of these metrics with the post-task questionnaire ratings of leadership and contribution, and used multiple linear regression to explain the variability of the leadership and contribution scores using these automated metrics. Our preliminary experiments show that using a combination of visual, non-verbal and verbal metrics, we can explain 65% and 63% of the leadership and contribution scores respectively. The metrics also could predict perceived group leaders and major contributors with 90% and 100% accuracy respectively [5].

6 CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we presented a multi-modal corpus for studying face-to-face group interactions. We used the Lunar Survival Task as our task-based collaborative discussion platform. As opposed to existing corpora on group meeting analysis that use frontal video cameras and wearable sensors, we propose the use of ceiling-mounted unobtrusive and identity-preserving depth sensors to capture the visual interactions.

We believe this dataset can be used by an interdisciplinary research community for the development of automated multimodal algorithms for human behavior analysis (e.g., VFOA, body pose, speech analysis), as well as by social psychology and organizational behavior researchers interested in group dynamics study. We presented our algorithms for head pose and VFOA estimation, and NLP techniques for opinion extraction and graphical summarization. We also presented preliminary analyses on perceived leadership and contribution using the automatically extracted audio-visual metrics. Our initial results show that even in the absence of rich frontal RGB data, we can derive significant levels of understanding of the evolving and emerging team patterns in face-to-face interactions. One of our current research directions using the UGI corpus is the development of an algorithm to automatically estimate arm and body pose, which can provide non-verbal correlates of several psychological variables like "trust", "liking" and "rapport" that affect group performance.

The preliminary analyses give us interesting findings on the effects of group composition and presence of women on task completion, group performance and perceived leadership. While each participant individually completed the item rankings, only 13 out of the 22 groups could come to a consensus on all 15 items in the stipulated time. Of the 13 groups that finished the task, 12 of them had at least 30% women (at least 1 woman in a group of 3, or 2 women in a group of 4 or 5). Of the 9 groups that did not finish, 8 fell short of this threshold. However, the highest performing groups had exactly 50% women and 50% men, and groups with all women performed only slightly better than groups with all men; thus, this

looks like a thresholding effect. This opens up new research questions about what women and men do differently that can provide different outcomes on the same task. Having a means to automatically compute audio-visual metrics enables us to study what verbal and non-verbal behaviors encourage participation, cohesion and contribution. Understanding these team processes can not only provide insights into human behavior in groups, but would also facilitate the development of active meeting facilitation systems that can help keep meetings on track and improve productivity.

ACKNOWLEDGMENTS

This work was supported by the NSF under award IIP-1631674 from the PFI:BIC program, by the NSF under cooperative agreement EEC-0812056, and by New York State under NYSTAR contract C090145. We also want to extend our thanks to Devavrat Jivani and Gyanendra Sharma for their help in setting up the sensors in the conference room.

REFERENCES

[1] Audacity. 2017. Audacity. http://www.audacityteam.org/. [Online; accessed 25-July-2018].
[2] Frank J Bernieri, Janet M Davis, Robert Rosenthal, and C Raymond Knee. 1994. Interactional synchrony and rapport: Measuring synchrony in displays devoid of sound and facial affect. *Personality and Social Psychology Bulletin* 20, 3 (1994), 303–311.
[3] Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. 2017. Multi-task learning of social psychology assessments and nonverbal features for automatic leadership identification. In *Proc. 19th Int. Conf. Multimodal Interaction*. ACM.
[4] Cigdem Beyan, Nicolò Carissimi, Francesca Capozzi, Sebastiano Vascon, Matteo Bustreo, Antonio Pierro, Cristina Becchio, and Vittorio Murino. 2016. Detecting emergent leader in a meeting environment using nonverbal visual features only. In *Proc. 18th ACM Int. Conf. Multimodal Interaction*. ACM.
[5] Indrani Bhattacharya, Michael Foley, Ni Zhang, Tongtao Zhang, Christine Ku, Cameron Mine, Heng Ji, Christoph Riedl, Brooke Foucault Welles, and Richard J Radke. 2018. A Multimodal-Sensor-Enabled Room for Unobtrusive Group Meeting Analysis. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 347–355.
[6] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. 2013. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image and Vision Computing* 31, 2 (2013), 203–221.
[7] Mckenzie Braley and Gabriel Murray. 2018. The Group Affect and Performance (GAP) Corpus. *Proceedings of Group Interaction Frontiers in Technology (GIFT 2018). ACM* (2018).
[8] Susanne Burger, Victoria MacLaren, and Hua Yu. 2002. The ISL meeting corpus: The impact of meeting type on speech style. In *INTERSPEECH*. Denver, CO.
[9] Nick Campbell, Toshiyuki Sadanobu, Masataka Imura, Naoto Iwahashi, Suzuki Noriko, and Damien Douxchamps. 2006. A multimedia database of meetings and informal interactions for tracking participant involvement and discourse flow. In *Proc. Int. Conf. Lang. Resources Evaluation*. Genoa, Italy.
[10] Tanzeem Choudhury and Alex Pentland. 2003. Sensing and Modeling Human Networks using the Sociometer. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers*. IEEE Computer Society, 216.
[11] Matthew A Cronin, Laurie R Weingart, and Gergana Todorova. 2011. Dynamics in groups: Are we there yet? *Academy of Management Annals* 5, 1 (2011), 571–612.
[12] Daniel Gatica-Perez. 2009. Automatic Nonverbal Analysis of Social Interaction in Small Groups: A Review. *Image Vision Comput.* 27, 12 (2009), 1775–1787.
[13] Theodoros Giannakopoulos and Aggelos Pikrakis. 2014. *Introduction to Audio Analysis: A MATLAB® Approach*. Academic Press.
[14] Jay Hall and Wilfred Harvey Watson. 1970. The effects of a normative intervention on group decision-making performance. *Human Relations* 23, 4 (1970), 299–317.
[15] Hayley Hung and Daniel Gatica-Perez. 2010. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Trans. Multimedia* 12, 6 (2010), 563–575.
[16] Hayley Hung, Yan Huang, Gerald Friedland, and Daniel Gatica-Perez. 2011. Estimating dominance in multi-party meetings using speaker diarization. *IEEE Trans. Audio, Speech, and Language Process.* 19, 4 (2011), 847–860.
[17] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, Vol. 1. I–I.
[18] Dineshbabu Jayagopi, Dairazalia Sanchez-Cortes, Kazuhiro Otsuka, Junji Yamato, and Daniel Gatica-Perez. 2012. Linking speaking and looking behavior patterns with group composition, perception, and performance. In *Proc. 14th ACM Int. Conf. Multimodal Interaction*. ACM.
[19] David W Johnson and Frank P Johnson. 1991. *Joining together: Group theory and group skills*. Prentice-Hall, Inc.
[20] Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. 2006. A corpus for studying addressing behaviour in multi-party dialogues. *Language Resources and Evaluation* 40, 1 (2006), 5–23.
[21] Jill Kickul and George Neuman. 2000. Emergent leadership behaviors: The function of personality and cognitive ability in determining teamwork performance and KSAs. *Journal of Business and Psychology* 15, 1 (2000), 27–51.
[22] Taemie Kim, Erin McFee, Daniel Olguin Olguin, Ben Waber, and Alex Pentland. 2012. Sociometric badges: Using sensor technology to capture new forms of collaboration. *Journal of Organizational Behavior* 33, 3 (2012), 412–427.
[23] Steve WJ Kozlowski. 2015. Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations. *Organizational Psychology Review* 5, 4 (2015), 270–299.
[24] Roger Th. A.J. Leenders, Noshir S Contractor, and Leslie A DeChurch. 2016. Once upon a time: Understanding team processes as relational event networks. *Organizational Psychology Review* 6, 1 (2016), 92–115.
[25] Nale Lehmann-Willenbrock, Hayley Hung, and Joann Keyton. 2017. New frontiers in analyzing dynamic group interactions: Bridging social and computer science. *Small Group Research* 48, 5 (2017), 519–531.
[26] Kurt Lewin. 1943. Psychology and the process of group living. *The Journal of Social Psychology* 17, 1 (1943), 113–131.
[27] Kurt Lewin. 1946. Action research and minority problems. *Journal of social issues* 2, 4 (1946), 34–46.
[28] Robert G Lord. 1977. Functional leadership behavior: Measurement and relation to social power and leadership perceptions. *Administ. Sci. Quart.* (1977), 114–133.
[29] Nadia Mana, Bruno Lepri, Paul Chippendale, Alessandro Cappelletti, Fabio Pianesi, Piergiorgio Svaizer, and Massimo Zancanaro. 2007. Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection. In *Proceedings of the 2007 workshop on Tagging, mining and retrieval of human related activity information*. ACM, 9–14.
[30] Wenxuan Mou, Hatice Gunes, and Ioannis Patras. 2016. Alone versus in-a-group: A comparative analysis of facial affect recognition. In *Proc. ACM Multimedia Conf.* ACM, 521–525.
[31] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting low rapport during natural interactions in small groups from non-Verbal behaviour. *arXiv preprint arXiv:1801.06055* (2018).
[32] Gabriel Murray, Hayley Hung, Joann Keyton, Catherine Lai, Nale Lehmann-Willenbrock, and Catharine Oertel. 2018. Group Interaction Frontiers in Technology. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 660–662.
[33] Catharine Oertel, Kenneth A Funes Mora, Samira Sheikhi, Jean-Marc Odobez, and Joakim Gustafson. 2014. Who will get the grant?: A multimodal corpus for the analysis of conversational behaviours in group interviews. In *Proc. Workshop Understanding Modeling Multiparty, Multimodal Interactions*. ACM.
[34] Daniel Olguın Olguın and Alex Sandy Pentland. 2007. Sociometric badges: State of the art and future applications. In *IEEE 11th Int. Symp. Wearable Comput.* Boston, MA.
[35] Kazuhiro Otsuka, Yoshinao Takemae, and Junji Yamato. 2005. A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances. In *Proc. 7th Int. Conf. Multimodal Interfaces*. ACM.
[36] Dairazalia Sanchez-Cortes, Oya Aran, and Daniel Gatica-Perez. 2011. An audio visual corpus for emergent leader analysis. In *Workshop Multimodal Corpora Mach. Learning: Taking Stock and Road Mapping the Future*. Alicante, Spain.
[37] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, and Phil Hall. 2017. English Conversational Telephone Speech Recognition by Humans and Machines. In *Proc. INTERSPEECH*.
[38] Stefan Scherer, Nadir Weibel, Louis-Philippe Morency, and Sharon Oviatt. 2012. Multimodal prediction of expertise and leadership in learning groups. In *Int. Workshop Multimodal Learning Analytics*.
[39] Thomas J. L. van Rompay, Dorette J. Vonk, and Marieke L. Fransen. 2009. The eye of the camera: Effects of security cameras on prosocial behavior. *Environment and Behavior* 41, 1 (2009), 60–74.
[40] Ni Zhang, Tongtao Zhang, Indrani Bhattacharya, Heng Ji, and Richard J. Radke. 2018. Visualizing Group Dynamics based on Multiparty Meeting Understanding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 96–101.