A Multimodal-Sensor-Enabled Room for Unobtrusive Group Meeting Analysis

Indrani Bhattacharya
Rensselaer Polytechnic Institute
Troy, New York
bhatti@rpi.edu

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

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

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

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

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

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

Group meetings can suffer from serious problems that undermine performance, including bias, "groupthink", fear of speaking, and unfocused discussion. To better understand these issues, propose interventions, and thus improve team performance, we need to study human dynamics in group meetings. However, this process currently heavily depends on manual coding and video cameras. Manual coding is tedious, inaccurate, and subjective, while active video cameras can affect the natural behavior of meeting participants. Here, we present a smart meeting room that combines microphones and unobtrusive ceiling-mounted Time-of-Flight (ToF) sensors to understand group dynamics in team meetings. We automatically process the multimodal sensor outputs with signal, image, and natural language processing algorithms to estimate participant head pose, visual focus of attention (VFOA), non-verbal speech patterns, and discussion content. We derive metrics from these automatic estimates and correlate them with user-reported rankings of emergent group leaders and major contributors to produce accurate predictors. We validate our algorithms and report results on a new dataset of lunar survival tasks of 36 individuals across 10 groups collected in the multimodal-sensor-enabled smart room.

ICMI '18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3243022

CCS CONCEPTS

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

KEYWORDS

Multimodal sensing; smart rooms; time-of-flight sensing; head pose estimation; natural language processing; meeting summarization; group meeting analysis

ACM Reference Format:
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 *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3242969.3243022

1 INTRODUCTION

Management studies report that tens of millions of meetings take place every day in the US, incurring a tremendous cost in terms of managers' and employees' precious time and salary [39]. Meetings are often inefficient, unfocused, and poorly documented. Any steps to make group meetings for complex, long-term projects more productive and easier to control would have immediate economic impact.

Automated group meeting facilitation can be passive or active. Passive meeting facilitation includes understanding participation and productivity shifts; when and why a meeting becomes unproductive; how factors like rapport, emergent team processes, mutual attentiveness, and coordination affect the productivity of a meeting; and how participants assume emergent leadership roles. Improved

measurement of emergent team processes is critical to advance the theory and understanding of collaborative decision making in groups [46]. A passive automated meeting facilitator could generate a retrospective summary of a meeting, as well as analyze the meeting to provide insights into the human interaction pattern that occurred.

On the other hand, an active automated meeting facilitator could aid in group decision making in real time. It could mediate the meeting by subtly reminding participants to talk more or less based on the real-time estimated speaking balance in the room. The room could additionally detect productivity shifts due to inattentiveness or lack of coordination and instigate a new line of thought. Furthermore, it could keep the meeting on track by correlating the agenda and the discussion schedule in real time.

A smart service system for passive and active meeting facilitation needs to automatically estimate participants' body and head poses, as well as record their speech. In existing work that analyzes group meetings, the locations of participants and their head poses and gaze directions are either manually annotated [37], which can be extremely time-consuming, or alternatively estimated using special wearable sensors [42], one or more cameras [10, 27], or front-facing Microsoft Kinects [40]. However, the presence of visible cameras can alter participants' natural behavior [51], and having unusual sensors directly in front of ones' face or in the line of sight may also inhibit natural individual behavior or group interactions. We posit that a room in which the participants are as unaware of being sensed as possible is best for studying natural group dynamics.

In this paper, we propose a multimodal-sensor-enabled room to facilitate group meetings, allowing the automatic extraction of a wide array of complementary metrics. The participants have natural conversations across a table with no sensors in their lines of sight while being recorded from above with ceiling-mounted, downward-pointed time-of-flight (ToF) distance sensors. We present a novel method to process the sensors' data to automatically estimate each participant's head pose and visual focus of attention. In parallel, the participants' microphone signals are fed through speech-to-text software and processed with natural language processing algorithms to automatically extract opinions and identify targets, resulting in play-by-play minutes of the meeting as it evolves. We tested the data extraction and multimodal signal processing algorithms on a new dataset of 36 individuals across 10 groups completing the lunar survival task in this sensor-instrumented smart room. We derive several spatial, non-verbal and verbal metrics from the different modalities and study their correlation with participants' post-task assessments of individuals' leadership and contribution, resulting in a linear regressor that accurately predicts perceived emergent leaders and perceived major contributors.

2 RELATED WORK

Studying human dynamics in face-to-face small group interactions is an active area of research. Perez [21] 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 range of topics in the automatic analysis of these social interactions includes interaction management (addressee, turn-taking), internal states (interest, other states), dominance (extroversion, dominance, locus of control) and roles (relationships). Murray [38] studied the relationship between the productivity of a meeting and linguistic and structural features. Lai et al. [33] showed how turn-taking patterns affect the perception of satisfaction and cohesion. Kim and Rudin [31] showed how analyzing local dialogue acts can predict when key decisions are being made in a group meeting.

However, much of this research heavily depends on human coding of events from recorded data. For example, Mathur et al. [37] developed a method for detecting interaction links between participants in a meeting using manually annotated video frames. The participants were asked to wear brightly-colored vests and personal audio recorders, and manual coding was used to localize each participant, record whether they were sitting or standing, and estimate whether they were speaking.

Automatic analysis of group interactions involves multimodal recording together with signal processing and computer vision techniques to derive various metrics for group dynamics analysis. Several multi-modal corpora have been designed for analysis of group meetings, using different combinations of modalities. These include the ICSI Meeting Corpus [26] (head-worn and table-top microphones), the ISL meeting corpus [14] (microphones), the AMI corpus [30] (video cameras and microphones), the ATR database [15] (small 360-degree camera surrounded by an array of high-quality directional microphones), the NTT corpus [41–43] (video cameras, microphones and wearable sensors), and the ELEA corpus [48] (close-talking mono-directional microphones, Kinects, and GoPro cameras).

Understanding the visual focus of attention (VFOA) of meeting participants is an important part of the automatic analysis of group meetings. The head pose (the head position and the head orientation defined by the pitch, yaw and roll angles) is often taken as a proxy for estimating the gaze direction. In existing literature, researchers mostly use front-facing cameras [5, 6, 10, 27, 36, 50], front-facing Kinects [40], or wearable sensors [42] for the automatic estimation of VFOA. However, video cameras and Kinects facing meeting participants is unnatural and could make people uncomfortable or inhibited, defeating the original purpose of studying natural human behavior in group interactions. In this paper, we propose a novel method for head pose and VFOA estimation in group meetings using ceiling-mounted, downward-facing Kinects. Mounting the Kinects on the ceiling makes them unobtrusive, well out of the fields of view of meeting participants. To the best of our knowledge, there is no work that employs ceiling-mounted Kinects for head pose and VFOA estimation, but as we show, our VFOA estimation accuracy is comparable to previous systems.

The correlation between non-verbal metrics (VFOA and speech signal based) and social-psychological group variables, such as perceived leadership, perceived dominance, and perceived extroversion was studied by Jayagopi et al. [27]. Beyan et al. studied the prediction of emergent leadership from non-verbal metrics [8, 10], and further investigated the problem of predicting the leadership style (autocratic or democratic) of an emergent leader [9]. In existing work on emergent leadership, researchers often employ personality trait-based questionnaires (e.g., agreeableness, conscientiousness, extroversion, neuroticism, openness) [27], such as NEO-FFI [29],

the General Leader Impression Scale (GLIS) [35], or manual annotation [8, 10]. In contrast, we do not use any personality-trait-based questionnaire or manual annotation for emergent leadership analysis because we are interested in who the group perceives as a leader. Such perception of the leader can be subjective and may vary from individual to individual. For example, someone who helps keep the discussion on track may be considered as a leader by some people, while someone who suggests the best answers to a problem may be considered as a leader by others. We leave this open to interpretation for the meeting participants and use a leadership score assigned by the participants as our metric for leadership. Perceptions, and in particular the convergence of perception, are important shared properties of the team that have bearings on team performance [32].

3 THE MULTIMODAL-SENSOR-ENABLED MEETING ROOM

Our experimental testbed is an $11' \times 28'$ conference room with two types of ceiling-mounted, downward-pointed Time-of-Flight (ToF) sensors [2]. These include 18 low-resolution IRMA Matrix ToF sensors, designed by Infrared Intelligent Systems (IRIS) and two higher-resolution Microsoft Kinect sensors, positioned over each side of the table, in order to capture seated individuals' head pose. Since the sensors are all embedded in the ceiling, they are outside participants' sight lines and there is no sense of being "watched". ToF sensors are advantageous compared to cameras in that (1) they return distance maps instead of images, enabling the direct creation of 3D point clouds of the environment, and (2) they are more robust to variations in the ambient lighting in the environment and the color/reflectiveness of the participants' clothing.

We use lapel microphones on each participant to record audio information. We removed noise from the audio signals in Audacity [4] and then performed automatic speaker identification using techniques described in [23]. 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. Accurate timestamps also allowed us to downsample the speaker identification information (collected at 48kHz) to the Kinect frame rate of 15fps. Thus, for a meeting with P participants, at each Kinect frame, we have a P-bit speaker label, where each bit denotes whether the participant is speaking or not. We extract several individual non-verbal metrics from the segmented speech, similar to [27]. These include speaking lengths, interruptions, speaking overlap, and backchannels (short responses like "uh-huh", "yes", "hmm", that are less than 2 seconds in duration, consistent with the definition in [27]). The first section of Table 1 lists these non-verbal metrics.

The recorded audio was transcribed to text using IBM Watson's Speech-to-Text API [49], which uses Long Short-Term Memory (LSTM) and Residual (ResNet) neural networks. The automatic transcriptions were further manually touched up to ensure the transcription was accurate and to compare algorithmic performance on raw vs. processed text.

Thus, the overall recorded multimodal data included the lower-resolution depth map from the 18 overhead IRMA Matrix ToF sensors (at 30 fps), the higher-resolution 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.

In order to synchronize the different modalities, each meeting discussed below started with a clap from a non-participant. Each of the lapel microphones, the two Kinects, the IRMA Matrix ToF sensors and the reference video camera recordings were synced together using the clap as the audio-visual cue for the start of the meeting. The VFOA of each of the meeting participants was manually annotated for 1 of the 10 meetings using the video camera recordings. This resulted in a dataset of approximately 45320 examples of ground-truthed VFOAs.

4 THE LUNAR SURVIVAL TASK DATASET

We recorded 36 individuals across 10 groups who completed the Lunar Survival Task [24] in the multimodal-sensor-enabled meeting room, which forms our current dataset for meeting understanding. 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. Each discussion lasts from 10–15 minutes, after which the participants are asked to complete a post-task questionnaire. In addition to questions relating to the age and gender of the participants, 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?
- To what extent did the following group members contribute to the discussion? [34]
- 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. [7]

The discussions were in English; based on self-reports, 40% of the participants were White, 46% Asian and 12% Hispanic/Latino. The ages of the participants ranged from 18 to 29 years and 40% of the participants were women.

5 HEAD POSE AND VFOA ESTIMATION USING TIME-OF-FLIGHT SENSORS

The general setup of each meeting is illustrated in Figure 1, which shows one frame from the reference camera view of a meeting, the low-resolution ToF depth map, and the corresponding location tracking and coarse body orientation estimation results.

The sparse ToF sensors are sufficient for occupant tracking [12, 28] and coarse body orientation estimation [11], but they do not provide enough information for head pose estimation. Hence,

Table 1: Automatically extracted non-verbal, visual, and verbal metrics.

Description	Symbol
Non-verbal Metrics (NV)	
Fraction of speaking length	SL
Fraction of silence	FS
Fraction of single person speaking	F1S
Fraction of two people speaking together	F2S
Fraction of three people speaking together	F3S
Fraction of four people speaking together	F4S
Fraction of successful interruptions	FSI
Fraction of unsuccessful interruptions	FUI
Fraction of backchannels	FBC
Fraction of speaking turns	FST
Visual Metrics (VZ)	
Attention received by a participant	ATR
Attention given by a participant	ATG
Attention Quotient (Ratio of ATR and ATG)	ATQ
Attention Center (fraction of time a participant is looked at by all other participants)	ATC
Attention Center of two people	ATC2
Attention Center of at least two people	ATCa2
Fraction of mutual gaze	FMG
Fraction of shared gaze	FSG
Verbal Metrics (VB)	
Fraction of times a participant acted as topic proposer	ITP
Fraction of times a participant acted as ranking proposer	RAP
Fraction of times a participant mentioned the decision ranking	DM
Relative rate at which a participant mentioned the decision ranking	FDM
Fraction of times a participant proposed the decision ranking for the first time	DMP
Fraction of times a participant summarized the decision ranking	DMS
TF-IDF informativeness score	TFIDF
Averaged informativeness	INFO
Fraction of number of sentences spoken by a participant	FOS

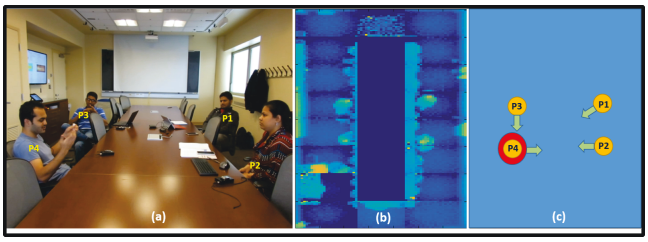

Figure 1: (a) Camera view, (b) Raw data from the sparse ToF sensors stitched to form a depth map of the room, (c) Results of the body orientation estimation algorithm: the red circle indicates the speaker, detected from the microphone recordings. The green arrows indicate the automatically estimated body orientations.

we use two ceiling-mounted Kinects for this purpose; Figure 2 illustrates an example elevation map from one of these Kinects, which contains the heads and upper torsos (with the chairs) of the seated individuals. While head pose estimation from depth sensors is an active area of research [13, 18, 44, 47], existing literature assumes that the head is viewed from the front, not from above, so different algorithms are required for our system. We now describe our framework for processing overhead depth data to estimate head pose and VFOA.

After thresholding based on the table height, the largest two connected components in Figure 2 correspond to the two human torsos. Segmenting the heads of people from the elevation map requires the determination of threshold values that are specific to each person. This value depends on several factors including the lengths of people's heads, their heights, and their pose (sitting straight/leaning over the table). Thus, it is difficult to determine a universal value that can be used for every person. In order to decide the per-person threshold value, we compute a histogram from the elevation values of the upper torso of each individual over 100 frames spread throughout the meeting duration, as illustrated in Figure 3. Here, the peak with elevation values around 1400mm corresponds to the head, and the valley around elevation value 1280mm corresponds to the neck and shoulder region. The rest of the histogram with lower elevation values corresponds to the portion of the body below the shoulders. Thus, detecting the minimum point in the first valley from the right in the histogram gives a threshold value for segmenting the head of an individual. Computing the histogram from 100 frames distributed over the entire meeting accounts for fluctuations due to movement of the body (leaning over the table, leaning back on the chair), the position of the hands with respect to the head (hands placed somewhere on the head/cheek), or noisy depth data. This dynamic and person-specific computation of the threshold produces a good head segmentation for the majority of the frames.

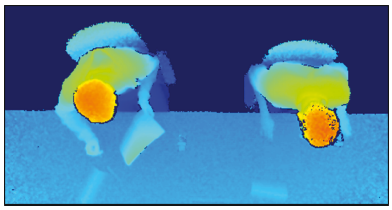

Figure 2: The depth map from one overhead Kinect sensor.

5.1 Ellipsoid Fitting to 3D Heads

The next step is to estimate 3D head orientations of each participant at each time instant. We compute 3D point clouds from the depth map of each of the two Kinects, and apply a rigid transformation that optimally aligns the two point clouds in the least-squares sense to build a combined 3D point cloud of the entire scene. Each segmented head is mapped from the 2D elevation map to the combined 3D point cloud. We now fit an ellipsoid to these 3D head points in each frame.

We parameterize a rotated 3D ellipsoid using 9 parameters $v = [a, b, c, d, e, f, g, h, i]$ as

$$aX^2 + bY^2 + cZ^2 + 2dXY + 2eXZ + 2fYZ + 2gX + 2hY + 2iZ = 1$$

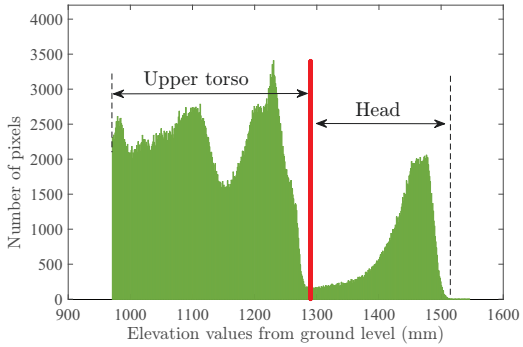

Figure 3: The histogram of elevation values of the upper torso of an individual over 100 frames. The red line is the estimated threshold value for head segmentation.

The set of N 3D head data points is used to build an $N\times9$ data matrix D in which each row has the form

$$D = [X_i^2, Y_i^2, Z_i^2, 2X_iY_i, 2X_iZ_i, 2Y_iZ_i, 2X_i, 2Y_i, 2Z_i]$$

The least-squared-error approximation of the unknowns is $v = (D^\top D)^{-1}D^\top 1_{N\times1}$. The center of the ellipsoid, the three radii, and the three axes directions can be computed from the vector v. We choose the ellipsoid axis that is closest to the unit vector pointing from the head center to the center of the table as the head pose. We found that several factors including noise, poor reflection from black hair, different hairstyles, stray hairs, headgear like caps/hoodies, hands on cheeks, and so on, mean that the segmented head and the corresponding 3D head points may not all be a good fit for an ellipsoid. Thus, we use the RANdom SAmple Consensus (RANSAC) algorithm [20] to get rid of outliers during the fitting process. We use a threshold inlier ratio of 0.8; that is at least 80% of the selected head points should be inlier points. Figure 4 illustrates the result of the ellipsoidal fit and head pose estimation for a sample participant.

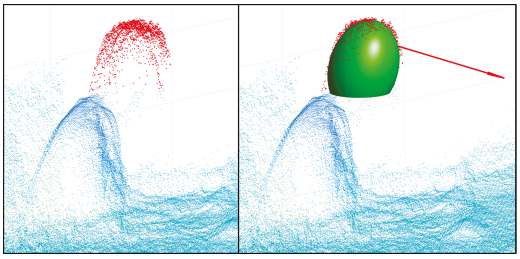

Figure 4: Head pose estimation for a sample participant. The left figure shows the 3D point cloud with the segmented 3D head points in red. The right figure shows the fitted ellipsoid with the head pose as a red arrow.

A good ellipsoid fit should have the 3D head points distributed uniformly all over the surface of the ellipsoid. To test the goodness of fit, we devised a metric, the *ellipsoid score*, defined as the sum of the chi-squared error and the distance between the mean of the 3D head points and the ellipsoid center. A small chi-squared distance means that the ellipsoid surface is close to the head points. A small distance between the ellipsoid center and the mean of the 3D head points ensures that the points are more or less uniformly distributed over the surface of the ellipsoid. Thus, a good fit will have a small ellipsoid score. We also define a confidence measure as the inverse of the ellipsoid score. After an initial pass of computation of head pose for all the Kinect frames of a meeting, we use the normalized confidence measure to smooth the head arrow directions using a weighted moving average technique. Figure 5 shows one frame of the 3D reconstructed point cloud and the fitted ellipsoids and head pose directions for all participants.

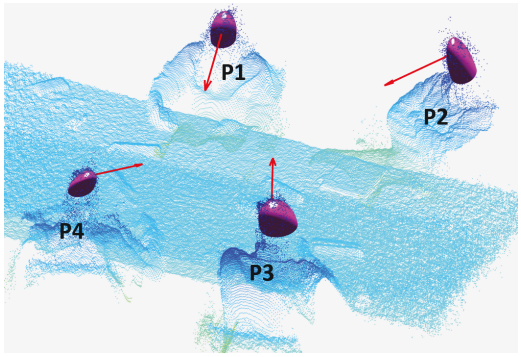

Figure 5: Ellipsoid fitting and head pose for all participants. Since not all points are good for ellipsoid fitting, we use RANSAC for removing outliers.

5.2 Estimating VFOA from computed head pose

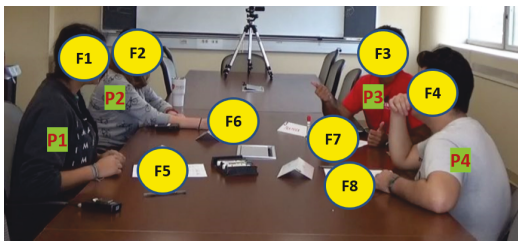

Figure 6: The VFOA target locations in our meeting scenario.

In our scenario, we posit that the VFOA target locations for a group meeting participant can either be one of the other participants, the piece of paper in front of her/him, or somewhere completely different. Figure 6 shows the labeled VFOA target locations for a meeting frame in our dataset. If a participant is not looking at any of these 8 locations, the VFOA target is taken as class 9. Thus, for example, the viable VFOAs for Person 1 are locations 2, 3, 4, 5, or 9. We used one meeting as our training data set and computed the probabilities of each person looking at another participant (either a speaker or non-speaker), the paper in front of them, or elsewhere. The results are listed in Table 2. We see that meeting participants spend approximately 50% of the time looking at the paper in front of them. This is because completing the task requires them to frequently refer to the list of items for comparative decision making, and to write down the ranking based on the group consensus for each item.

Table 2: Annotated VFOA probabilities for a participant based on training data.

VFOA → Participant is ... ↓	speaking participant	paper	other participant	elsewhere
speaking	0.14	0.50	0.35	0.01
not speaking (someone else is speaking)	0.33	0.47	0.19	0.01
not speaking (nobody else is speaking)	0	0.49	0.50	0.01

Our VFOA estimator is a Bayes classifier that uses both the depth information from the Kinect sensors and the synchronized speaker identification information from the microphone recordings. The likelihood term is computed from the head pose measurements while the prior term is computed dynamically at each frame based on the speaker information.

At each Kinect frame, we have the head pose of each participant, which is represented as a unit vector from the center of the fitted ellipsoid. We also have the 3D coordinates of all 8 target locations. We then compute the angle between the head pose and each of the target locations. The inverses of these angles, normalized to sum to 1, become our likelihoods. Since the audio and Kinect frames are synchronized, we can find the speaker/s at each Kinect time instant by speaker segmentation as described in Section 3. We compute the prior probability of each participant looking at any of the VFOA target locations based on her/his role as a speaker or a listener using the computed probabilities from the training data in Table 2. Multiplying the likelihood with the prior gives the posterior probability distribution, and the target that has the highest posterior probability is considered as the VFOA of the participant. Since the VFOAs of participants do not change abruptly, we post-process the estimated target locations with a median filter of window size 15 frames.

We evaluated the performance of our VFOA estimation algorithm on 1 manually annotated meeting of 13 minutes duration with 4 participants. The total number of frames was $11330 \times 4 = 45320$. The accuracy of our algorithm on this sample was **48.35%**. We note that researchers have reported comparable accuracy (42%) for VFOA estimation in similar group meeting settings with the same VFOA target locations using front-facing video cameras [27], while our sensor configuration is much less obtrusive. Recent work [9] on the ELEA corpus (which uses front-facing video cameras and Kinects) reports higher VFOA accuracies in the 80% range, but the target locations are less complex and not directly comparable to our setup.

After VFOA estimation, we further derive various visual metrics, such as the attention received and attention given by each participant, their ratio (attention quotient), and so on, as itemized in the middle section of Table 1. We demonstrate in the correlation and regression analyses in Section 7 that our VFOA estimation accuracy is sufficient to aid in accurately predicting perceived group leaders and contributors. A short video clip illustrating the head pose and VFOA estimation on a meeting segment can be viewed at https://youtu.be/ki5UJPSOZdE.

6 VERBAL SPEECH UNDERSTANDING

We now discuss verbal metrics obtained from automatic natural language processing algorithms applied to each microphone channel. Our goal is to extract measurements relevant to the participants' leadership or influence within the group. In particular, we consider how participants' opinions change throughout the conversation, and the amount and importance of information conveyed to the group by each speaker. The result is a set of "structured minutes" of the meeting that dynamically plays back the conversation, automatically highlighting each of the 15 lunar task supply items when they are discussed, along with the participants' opinions of their ranking. This refined "play-by-play" is represented internally as a bipartite graph [25] as described below. A snapshot of the automatically generated graphical summary is illustrated in Figure 7.

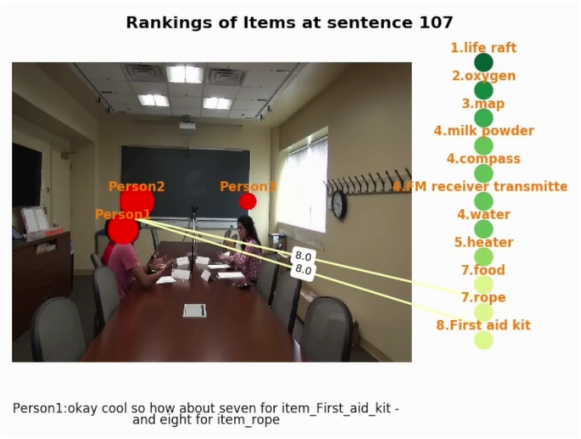

Figure 7: A snapshot of the graphical summarization of a meeting. The areas of the red circles are proportional to the cumulative speaking time of each participant. Straight lines connect the participants with their proposed items and item rankings.

6.1 Opinion extraction and target identification

In the lunar survival task discussion, participants can express their opinions of item ranks in several ways, including

(1) Explicitly ranking an item (e.g., "*The oxygen tank should be first.*")
(2) Agreement or disagreement (e.g., "*Yeah, I agree.*")
(3) Comparison of the items by relative ranking (e.g., "*Water should be higher than the compass.*")

In the first case of a participant proposing an item ranking, we use the Stanford CoreNLP Named Entity Recognizer (NER) annotator to extract the NUMBERs and ORDINALs mentioned in the discussion [19]. We also eliminated numbers beyond 15 and numbers that are parts of pronouns like "this one". We address the second case by observing that people typically express agreement/disagreement with the person who talked immediately before them [1]. For agreement, we assume the current speaker accepts

the previous speaker's stated opinion. For disagreement, we noted that people typically express their own opinion about a contested item immediately after disagreement (similar to the first case). We currently do not capture the third case of relative rankings, since no definitive ranking can be extracted from such a statement, and the objects are typically discussed at specific ranks later.

Identifying the target (i.e., object) of a ranking has two steps. First, we identify the term in a sentence that refers to the same item by a dictionary of synonyms collected from the observation. We then use a rule-based algorithm to link the rank closest to the item as the extracted pair.

To capture the instantaneous state of the discussion, we use an undirected bipartite graph $G = (U \cup V, E)$, where the vertices U represent the participants, the vertices V represent the items, and the edges E represent user rankings of objects. Each edge has a weight corresponding to the stated ranking. As the meeting continues, the bipartite graph is dynamically built up, allowing us to view the current opinions of items that have been discussed so far and to observe the opinion changes throughout the conversation. A short video clip illustrating the the graphical summarization of the meeting can be viewed at https://youtu.be/9z8OfmEtIuw.

6.2 Derived verbal metrics

Finally, we extract two categories of verbal metrics for the purpose of leadership modeling, as summarized at the bottom of Table 1.

6.2.1 Metrics related to the mentions of items/ranking-item pairs. To reach group consensus, participants may change their opinions based on the opinions expressed by other participants in the conversation, which is an influential process that could affect the leadership scores of these participants. The opinion extraction and target identification allows us to easily extract metrics related to this opinion change. In particular, these include the number of times each participant proposes an item or ranking for the first time, and the number of times a participant's expressed opinion agrees with the final group consensus, which we call decisive mentions. To explore whether perceived leadership is more related to proposing items/ranks or summarizing the group consensus, we count the number of decisive mentions a participant makes in the role of proposer and summarizer.

6.2.2 Metrics related to the efficiency of conveying information. We hypothesize that participants who make more informative utterances are perceived by the group as leaders. In the lunar survival scenario, we observed that emergent leaders often actively mentioned items (e.g., to propose a ranking, or to argue or defend against opinions on item usage), while the other participants made fewer explicit item mentions and their utterances usually consisted of expressions of general agreement or disagreement. To reflect this observation, we computed *TF-IDF* (term frequency-inverse document frequency) scores [45].

7 ANALYSIS OF LEADERSHIP AND CONTRIBUTION

Each meeting was followed by a post-task questionnaire as described in Section 4. Using the questionnaire, we define two target variables: perceived leadership and perceived contribution. Since each group member rates the leadership and contribution of all other group members on a 5-point scale, we compute an individual's perceived leadership score as the average of all leadership scores they receive from their group. The perceived contribution score is defined similarly.

We computed the Pearson Correlation Coefficient (ρ) to understand the correlation between the metrics and the target variables. We then regressed each target variable against the non-verbal audio, verbal, and visual metrics for each of the 36 participants to test the predictive power of the metrics that we compute, and to determine which metrics were most salient. First, single variable regressions were performed to determine the relationship between each of the metrics and the target variables.

Verbal metrics were overall the best predictors of leadership and contribution, as illustrated in Figure 8. DM correlated most strongly with leadership ($\rho = 0.46$, $p = 0.005$), since the leader is typically the person driving the discussion and working to finish the task. Leaders also propose topics, propose ranks, and summarized the decision ranking at significantly higher rates than other group members (all $p < 0.05$), yielding further evidence that leaders were driving the discussion. The relationship between the verbal metrics and perceived contribution was qualitatively similar. Even though FDM was most strongly correlated with perceived contribution ($\rho = 0.52$, $p < 0.001$), DM was also very strongly correlated ($\rho = 0.51$, $p < 0.001$), as were RAP ($\rho = 0.44$, $p < 0.001$) and DMS ($\rho = 0.40$, $p < 0.001$). However, ITP was not strongly correlated with perceived contribution. Thus top contributors were not likely to propose new discussion topics, in contrast with leaders.

Nonverbal metrics were much less salient. Among the 10 nonverbal metrics we computed, none were significantly associated with either perceived leadership or contribution. FSI had the strongest correlation with leadership ($\rho = 0.28$, $p = 0.1$) while FST had the most correlation with contribution ($\rho = 0.26$, $p = 0.1$). This can be interpreted in the sense that an individual who is able to successfully interrupt another person is listened to by the other participants and given importance as a leader.

Within the visual metrics, leadership was most significantly correlated with ATQ ($\rho = 0.37$, $p = 0.07$). Leaders tend to receive more visual attention than they give, though the association is just above the 0.05 level of significance. This is in line with observations made in [22], where the authors report that emergent leaders are looked at more often than other participants. Interestingly, the visual metric that most strongly correlates with perceived contribution is ATG, but the relationship is negative ($\rho = -0.44$, $p = 0.007$), suggesting that high level contributors tend to look at peers in the discussion less often than others.

We used multiple linear regression to investigate the capability of the extracted metrics to predict the post-task questionnaire variables of leadership and contribution. From the experiments, we found that a combination of all the visual, non-verbal and verbal metrics can explain 65% ($F = 1.39$, $p = 0.26$) of the variability of the leadership scores. Similarly, all the metrics combined can explain 63% ($F = 1.27$, $p = 0.32$) of the variability of the contribution scores, as shown in Figure 8.

The leadership scores and the contribution scores have a correlation coefficient of 0.63. Therefore, we can say that groups do not necessarily choose the participant who contributed the most as the leader, although the two variables are strongly correlated. Finally,

we used the linear regression coefficients to predict the leadership and contribution scores for each participant. Since the actual leadership scores are quantized, we also quantize the predicted scores to the nearest actual bin and find the participant(s) with the highest scores. An actual perceived group leader is the participant(s) with the highest received leadership score. Similarly, a predicted group leader is the participant(s) with the highest predicted quantized leadership score. We found that combining the visual, non-verbal and verbal metrics, we were capable of predicting the perceived emergent leader with an accuracy of **90%** (i.e., for 9 of the 10 meetings). Using a similar method for contribution, we found that we could predict the major contributor with a **100%** accuracy with the verbal metrics alone. This result is promising and shows that even without front-facing video cameras, we can analyze group meetings in terms of leadership and contribution to a plausible level of accuracy.

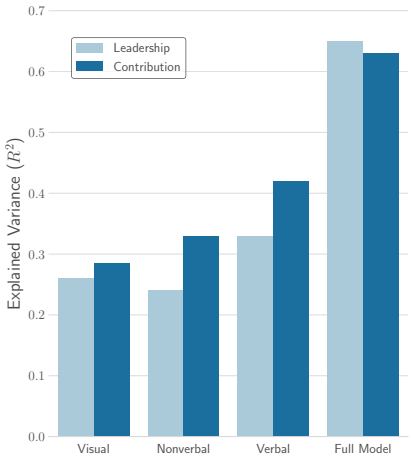

Figure 8: Comparison of 4 regression models on the two dependent variables, perceived leadership and perceived contribution to discussion. The first three models (visual, nonverbal, verbal) take into account all variables in each group, as specified in Table 1. The full model combines all variables from the first 3 models.

8 CONCLUSIONS AND FUTURE WORK

We presented a multimodal-sensor-enabled smart room that does not use video cameras or obtrusive sensing techniques that might make participants feel "watched". Our initial analysis of the lunar survival task experiments in this room provides good results for passive meeting understanding. We can predict group leaders and major contributors with 90% and 100% accuracy respectively, using the automatically extracted metrics. In future work, we plan to make this room an active meeting facilitator, subtly stepping in during a meeting to manage the participants and agenda.

One promising avenue for improved analysis involves the automated identification of the gender of meeting participants. Prior research has shown significant differences in the way that women and men interact and are perceived in team meeting environments. Indeed, controlling for gender in our regression model allows us to reduce the degrees of freedom of the model from 20 down to 13, while improving the R^2 and F-test value. For the perceived leadership regression, changing the model in this way yields an $R^2 = 0.686$, $F = 3.69$, p = 0.003. For the perceived contribution regression, changing the model in this way yields an $R^2 = 0.667$, $F = 3.01$, p = 0.011. For active meeting facilitation, we can infer gender from the recorded speech signal.

With regard to the ToF sensing modality, we want to modify our VFOA estimation algorithm to include supervised learning techniques with more contextual cues, which should improve the accuracy. We also want to implement more sophisticated supervised learning techniques for predicting leadership scores and styles of leadership by combining other metrics involving coarse body pose extracted from the lower-resolution ToF sensors, and body and head activities, as discussed in [9].

Currently, we are using individual lapel microphones for each participant. However, our ultimate aim is to have a smart room that is completely unobtrusive and does not require the participants to wear any specific sensors. We are working towards integrating custom 16-channel ambisonic (spherical) microphones [16, 17] into the smart room. The 16 channels can be combined differently to point at each of the instantaneous participant locations obtained by the ToF tracking system, allowing us to more clearly understand the focus of attention of participants in the meeting. A source segregation model can be used after extracting each auditory signal source using the beam-forming capabilities, which will further improve the signal-to-noise ratio. A better audio signal will also result in a cleaner automatic transcription process, resulting in less manual annotation effort.

The transcripts and derived bipartite graph summary provide rich information about what is being spoken. A fusion of natural language processing of the generated transcripts together with estimated VFOA can help understand participation shifts, i.e., how each participant changes role from being a speaker to an addressee to an unaddressed recipient.

We intend to record more lunar task experiments in the sensor-enabled smart room, to perform deeper statistical analysis to further correlate derived metrics with participant opinions, and design further studies to investigate team dynamics. We also plan on conducting further experiments on participants' perception of front-facing vs. ceiling-mounted sensors, focusing on intrusiveness/naturalness. While a tabletop device with more advanced sensing techniques can provide a richer dataset for analysis, it would not be suitable for situations where people can move about freely, such as cocktail parties or poster sessions. One direction for future research is to explore the use of ceiling-mounted ToF sensors to study social interaction patterns for free-standing conversations, similar to [3].

9 ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. IIP-1631674 and by a Northeastern University Tier 1 Seed Grant. We are grateful to the NSF for supporting student travel to attend ICMI 2018 under Grant No. IIS-1829325. We also extend our thanks to Devavrat Jivani and Gyanendra Sharma.

REFERENCES

[1] Amjad Abu-Jbara, Mona Diab, Pradeep Dasigi, and Dragomir Radev. 2012. Subgroup detection in ideological discussions. In *Proc. 50th Annu. Meeting Assoc.*

Comput. Linguistics: Long Papers-Volume 1. Assoc. Comput. Linguistics, 399–409.
[2] S. Afshari, T. K. Woodstock, M. H. T. Imam, S. Mishra, A. C. Sanderson, and R. J. Radke. 2015. The Smart Conference Room: An integrated system testbed for efficient, occupancy-aware lighting control. In *ACM Int. Conf. Embedded Syst. Energy-Efficient Built Environments*.
[3] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. 2016. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 8 (2016), 1707–1720.
[4] Audacity. 2017. Audacity. http://www.audacityteam.org/. [Online; accessed 25-July-2018].
[5] Sileye O Ba and Jean-Marc Odobez. 2009. Recognizing visual focus of attention from head pose in natural meetings. *IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 1 (2009), 16–33.
[6] Sileye O Ba and Jean-Marc Odobez. 2011. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Trans. Pattern Anal. and Mach. Intell.* 33, 1 (2011), 101–116.
[7] 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.
[8] 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*.
[9] Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. 2018. Prediction of the leadership style of an emergent leader using audio and visual nonverbal features. *IEEE Trans. Multimedia* 20, 2 (2018), 441–456.
[10] 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.
[11] Indrani Bhattacharya, Noam Eshed, and Richard J Radke. 2017. Privacy-preserving understanding of human body orientation for smart meetings. In *Int. Conf. Comput. Vision Pattern Regognition Workshops*. IEEE.
[12] Indrani Bhattacharya and Richard J Radke. 2016. Arrays of single pixel time-of-flight sensors for privacy preserving tracking and coarse pose estimation. In *Proc. Winter Conf. Appl. Comput. Vision*. IEEE.
[13] Guido Borghi, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara. 2017. Poseidon: Face-from-depth for driver pose estimation. In *Int. Conf. Computer Vision Pattern Recognition*. IEEE.
[14] Susanne Burger, Victoria MacLaren, and Hua Yu. 2002. The ISL meeting corpus: The impact of meeting type on speech style. In *INTERSPEECH*. Denver, CO.
[15] 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.
[16] Samuel Clapp, Anne Guthrie, Jonas Braasch, and Ning Xiang. 2013. Three-dimensional spatial analysis of concert and recital halls with a spherical microphone array. In *ASA Proc. Meetings Acoust.* Montreal, Canada.
[17] Samuel Clapp, Anne E Guthrie, Jonas Braasch, and Ning Xiang. 2013. Headphone- and loudspeaker-based concert hall auralizations and their effects on listeners' Judgments. *The J. Acoust. Soc. America* 134, 5 (2013), 3969–3969.
[18] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. 2013. Random forests for real time 3D face analysis. *Int. J. Comput. Vision* 101, 3 (2013), 437–458.
[19] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics*. Assoc. Comput. Linguistics, 363–370.
[20] Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395.
[21] Daniel Gatica-Perez. 2009. Automatic nonverbal analysis of social interaction in small Groups: A review. *Image Vision Comput.* 27, 12 (2009), 1775–1787.
[22] Fabiola H Gerpott, Nale Lehmann-Willenbrock, Jeroen D Silvis, and Mark Van Vugt. 2017. In the eye of the beholder? An eye-tracking experiment on emergent leadership in team interactions. *The Leadership Quarterly* (2017).
[23] Theodoros Giannakopoulos. 2009. A method for silence removal and segmentation of speech signals, implemented in Matlab. *University of Athens* (2009).
[24] 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.
[25] Xiangnan He, Ming Gao, Min-Yen Kan, and Dingxian Wang. 2017. Birank: Towards ranking on bipartite graphs. *Trans. Knowl. Data Eng.* 29, 1 (2017), 57–71.
[26] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The ICSI meeting corpus. In *Int. Conf. Acoust., Speech, and Signal Process.*
[27] 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.
[28] Li Jia and Richard J Radke. 2014. Using Time-of-Flight measurements for privacy-preserving tracking in a smart room. *IEEE Trans. Ind. Informatics* 10, 1 (2014), 689–696.
[29] Oliver P John and Sanjay Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook Personality: Theory and Research* 2, 1999 (1999), 102–138.
[30] 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.
[31] Been Kim and Cynthia Rudin. 2014. Learning about meetings. *Data Mining Knowl. Discovery* 28, 5-6 (2014), 1134–1157.
[32] S. W. Kozlowski. 2015. Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations. *Organizational Psychology Review* 5, 4 (2015), 270–299.
[33] Catherine Lai, Jean Carletta, and Steve Renals. 2013. Modelling participant affect in meetings with turn-taking features. In *Proc. Workshop Affective Social Speech Signals*.
[34] Robert G Lord. 1977. Functional leadership behavior: Measurement and relation to social power and leadership perceptions. *Administ. Sci. Quart.* (1977), 114–133.
[35] Robert G Lord, Roseanne J Foti, and Christy L De Vader. 1984. A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational Behavior and Human Performance* 34, 3 (1984), 343–378.
[36] Benoît Massé, Silèye Ba, and Radu Horaud. 2017. Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE Trans. Pattern Anal. Mach. Intell.* (2017).
[37] Shobhit Mathur, Marshall Scott Poole, Feniosky Pena-Mora, Mark Hasegawa-Johnson, and Noshir Contractor. 2012. Detecting interaction links in a collaborating group using manually annotated data. *Social Networks* 34, 4 (2012), 515–526.
[38] Gabriel Murray. 2014. Learning how productive and unproductive meetings differ. In *Canadian Conf. Artificial Intell.* Springer.
[39] Jay F Nunamaker Jr, Robert O Briggs, Daniel D Mittleman, Douglas R Vogel, and Balthazard A Pierre. 1996. Lessons from a dozen years of group support systems research: A discussion of lab and field findings. *J. Manage. Inform. Syst.* 13, 3 (1996), 163–207.
[40] 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.
[41] Kazuhiro Otsuka, Hiroshi Sawada, and Junji Yamato. 2007. Automatic inference of cross-Modal nonverbal interactions in multiparty conversations: Who responds to whom, when, and how? From gaze, head gestures, and utterances. In *Proc. Int. Conf. Multimodal Interfaces*. ACM, Aichi, Japan.
[42] 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.
[43] Kazuhiro Otsuka, Junji Yamato, Yoshinao Takemae, and Hiroshi Murase. 2006. Conversation scene analysis with dynamic Bayesian Network based on visual head tracking. In *Proc. Int. Conf. Multimedia and Expo*. IEEE, Toronto, ON, Canada.
[44] Pashalis Padeleris, Xenophon Zabulis, and Antonis A Argyros. 2012. Head pose estimation on depth data based on particle swarm optimization. In *Comput. Vision and Pattern Recognition Workshops*. IEEE.
[45] Jason DM Rennie and Tommi Jaakkola. 2005. Using term informativeness for named entity detection. In *Proc. 28th Annu. Int. SIGIR Conf. Research Develop. Inform. Retrieval*. ACM.
[46] Christoph Riedl and Anita Williams Woolley. 2017. Teams vs. crowds: A field test of the relative contribution of incentives, member ability, and emergent collaboration to crowd-based problem solving performance. *Academy Manage. Discoveries* 3, 4 (2017), 382–403.
[47] Anwar Saeed and Ayoub Al-Hamadi. 2015. Boosted human head pose estimation using Kinect camera. In *Int. Conf. Image Process.* IEEE.
[48] 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.
[49] 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*.
[50] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. 2002. Modeling focus of attention for meeting indexing based on multiple cues. *Trans. Neural Netw.* 13, 4 (2002), 928–938.
[51] 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.