Privacy-Preserving Understanding of Human Body Orientation for Smart Meetings

Indrani Bhattacharya, Noam Eshed, and Richard J. Radke
Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
bhatti@rpi.edu, eshedn@rpi.edu, rjradke@ecse.rpi.edu

Abstract

We present a method for estimating the body orientation of seated people in a smart room by fusing low-resolution range information collected from downward pointed time-of-flight (ToF) sensors with synchronized speaker identification information from microphone recordings. The ToF sensors preserve the privacy of the occupants in that they only return the range to a small set of hit points. We propose a Bayesian estimation algorithm for the quantized body orientations in which the likelihood term is based on the observed ToF data and the prior term is based on the occupants' locations and current speakers. We evaluate our algorithm in real meeting scenarios and show that it is possible to accurately estimate seated human orientation even with very low-resolution systems.

1. Introduction

Millions of meetings take place every day, resulting in a huge expenditure in terms of human energy, time and money [19]. However, meetings are often inefficient, unfocused, and poorly documented. A multi-modal sensor enabled physical environment with advanced cognitive computing capabilities could assist in making group meetings for long-term, complex projects more productive and easier to control, resulting in an immediate economic impact. Future smart service systems will be able to distinguish and clearly isolate the speech of several people talking at the same time, learn and remember the context of previous meetings, summarize what happened in the meeting, analyze participation shifts and meeting productivity, and ultimately contribute in real time to facilitate group decision making. Towards this end, the room in which the meeting occurs has to be smart enough to understand where people are, what their poses are, and in what direction their bodies are oriented. A natural solution is to use video cameras to track people and estimate their head and body poses. However, from a social perspective, meeting participants could feel uncomfortable, self-conscious, or inhibited in the presence of active video cameras.

In this paper, we present a method for classifying the seated orientation of participants in a meeting into one of eight quantized direction bins, as illustrated in Figure 1. Our Bayesian estimation algorithm uses two orthogonal modalities. The first is range (distance) information collected from an array of ceiling-mounted, downward-pointed time-of-flight (ToF) sensors. The ToF sensors produce a sparse range map of the room by analyzing the phase difference between emitted and reflected infrared signals. People appear as untextured blobs in the output of the ToF sensors, which makes them substantially more privacy-preserving than video cameras. The second modality is non-verbal audio information recorded from individual lapel microphones carried by each meeting participant. That is, we only determine which participants are speaking at each instant, not the words that are said. Figure 2 illustrates a frame of reference video with the corresponding recorded ToF and speaker identification.

The body orientation classification algorithm uses the ToF depth map of each person blob to compute the likelihood of each orientation class, based on a compressed sensing approach applied to examples of labeled training data. The prior probability distribution is computed dynamically at each frame for every person by analyzing the audio and ToF data to determine the active speakers and relative positions of the participants. The algorithm works with 80.8% accuracy to exactly classify seated orientations and with 98% accuracy to classify orientations with an error of 1 orientation class (± 45°).

2. Related Work

Organizational and social psychologists leverage principled probabilistic models for analyzing team dynamics, but such methods heavily depend on human coding of events from observed recorded data. For example, Mathur

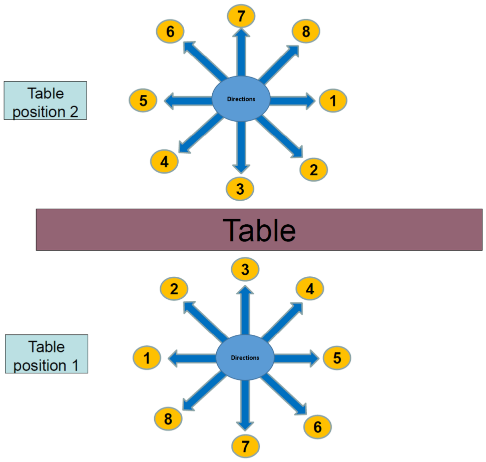

Figure 1. The 8 different orientation directions relative to the table.

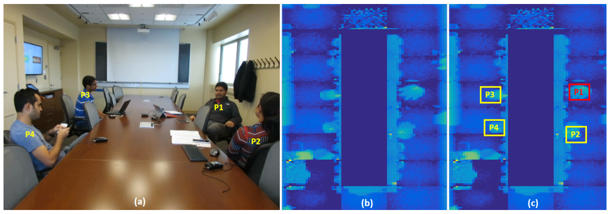

Figure 2. (a) Camera view, (b) Raw data from the ToF sensors stitched to form a depth map of the room, (c) Occupancy tracking using the ToF data.

et al. [18] developed a method for detecting interaction links between participants in a meeting using manually annotated video frames. The participants were required to wear brightly-colored vests and personal audio recorders, and manual coding was used to localize each participant in 78cm-wide cells, record whether they were sitting or standing, and estimate whether they were speaking. One frame for every 10 seconds of each test video was annotated at an average speed of 73 seconds per frame. Clearly, effective automated methods would be a boon to such social science analysis.

The analysis of non-verbal cues including location, head and body pose, gaze direction, hand gestures, speaker segmentation and meeting contextual cues are important for the automatic analysis of social interaction. Perez [10] 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). For example, Jovanovic et al. [16] worked on the problem of addressee identification using manually annotated data. The problem of estimating the visual focus of attention of participants from non-verbal cues is also an active area of research [26, 3, 11, 22, 21], which can be used to detect emergent leaders [5].

Several multimodal corpora have been designed for analysis of group meetings, with different combinations of modalities. These include the ICSI Meeting Corpus [13] (head-worn and table-top microphones), the ISI meeting corpus [6] (microphones), the AMI corpus [17] (video cameras and microphones), the ATR database [7] (small 360-degree camera surrounded by an array of high-quality directional microphones), the NTT corpus [24, 23, 22] (video cameras, microphones and wearable sensors), and the ELEA corpus [25] (close-talking mono-directional microphones, Windows Kinect and GoPro cameras).

In most of these studies, the locations of participants, head poses and gaze directions were either manually annotated [18, 16] or estimated using special wearable sensors [23], one or more cameras [26, 14, 5], or the Kinect [20]. The manual coding suffered from non-trivial inaccuracies in each type of measurement. Using cameras, separate Kinects for individual participants, or wearable sensors for measuring head pose is obtrusive and generally makes the participants uncomfortable. In contrast, in this paper, we present a system that automatically tracks participants and estimates their seated orientations, without the use of any video cameras or Kinects. To the best of our knowledge, there is no work that employs ceiling-mounted, sparse ToF sensors for understanding the orientation of participants in a group meeting, although such arrays of sensors are much more likely to integrate naturally into future building systems.

The likelihood for a test orientation image to belong to a particular class is computed based on the assumption that the test image approximately lies in the linear span of training samples from the same class. This idea is inspired from Wright et al. [27], which uses sparse representation to solve the problem of face recognition. The choice of features is less critical if the sparse representation is properly computed. We leverage the observation that participants in a meeting generally face the current speaker in order to compute the prior probability distribution of the orientation classes. We use a Bayesian estimation algorithm that combines likelihood and prior terms to automatically detect the seated body orientations of the participants.

3. Problem Statement and Dataset

3.1. Problem Statement

Given a meeting scenario, our task is to estimate the body orientation of all seated individuals at each ToF frame. We consider eight different orientation classes, which are defined by the participant location with respect to the table, as illustrated in Figure 1. In either of the two table positions,

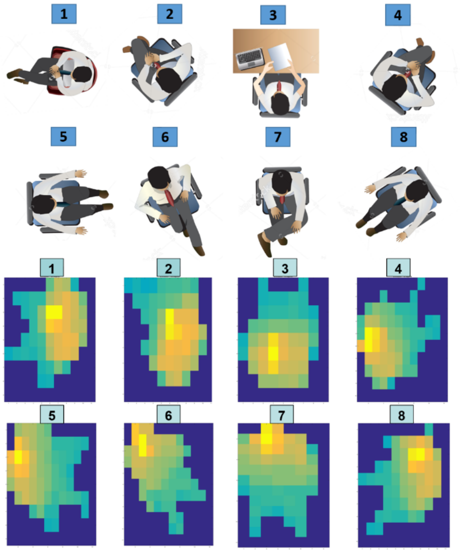

Figure 3. Top-down views of orientation classes and corresponding actual ToF images.

orientation direction 3 is towards the table. The top-down views of the eight orientations and corresponding representative ToF images for an individual are shown in Figure 3.

3.2. Dataset

The dataset for our study consists of a meeting with 4 participants, debating on a topic. The duration of the meeting was around 12 minutes. The participants were spontaneous during the discussions and nothing was scripted.

The meeting was conducted in an $11' \times 28'$ conference room with 18 ceiling-mounted IRMA Matrix time-of-flight sensors, designed by Infrared Intelligent Systems (IRIS) [1]. These sensors work on the principle of the time-of-flight of light, analyzing the phase difference between the emitted and reflected infrared signals to estimate the relative distance of an object from the sensor. The resolution of each sensor is 25×20 pixels. The depth map obtained from the sensor array is thus extremely low-resolution. Individuals are continuously tracked in the room and their coarse poses (sitting/standing) are determined from the output of the ToF sensors using blob tracking and height thresholding [15]. The ToF sensors collect data at approximately 9 frames per second (fps). Each participant also wore a lapel microphone. The ToF and microphone data were recorded and served as inputs to our algorithm. A reference video camera was also used to record the meeting proceedings. The video camera data was not used for any algorithm development and was only used for ground truth determination.

For the purpose of training the orientation classifier, we also conducted a separate non-meeting recording in which two different individuals sat in different parts of the room in each of the 8 different orientations. This resulted in 900 images of each class corresponding to Table Position 1 and 450 images of each class corresponding to Table Position 2. We will explain in Section 5 how we use these images to generate training datasets for our classifier.

4. Proposed Method

4.1. Pre-processing the data

At the start of the meeting, one of the participants waved his hand and verbally indicated the start of the meeting. The audio and reference video recordings were first synchronized in Audacity [2]. The reference video and ToF data were then synchronized using the waved hand. Time stamps on the ToF data were used to find correspondences between ToF frames (collected at 9 fps) and reference video frames (collected at 30 fps) and to appropriately downsample the video data.

We removed noise from the aligned audio data in Audacity and then performed speaker identification using techniques described in [12]. 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. Accurate timestamps also allowed us to downsample the speaker identification information (collected at 48kHz) to the ToF frame rate of 9 fps. Thus at each ToF frame, we have a 4-bit speaker label. There can be more than one speaker at a particular time or there can be no speaker at all. For example, if the speech label is $[0, 0, 0, 1]$, the speaker is P4, while if the speech label is $[0, 0, 1, 1]$, both P3 and P4 are speaking simultaneously.

4.2. Manual annotation

The true body orientations of all the participants in the meeting were manually annotated from the video camera recordings. To account for variability between different human annotators, we randomly selected 300 frames and annotated the body orientation for all the participants using two different annotators. The annotator agreement on human orientation was 92%, and always differed by 1 class in cases when the annotation was not in agreement.

4.3. Body orientation estimation

Our body orientation estimation algorithm uses a Bayes classifier applied to measurements from the two modalities. The likelihood term is computed by applying compressed sensing techniques to the ToF depth map. The prior term is calculated dynamically at each frame after extracting the speaker identity from the audio, combined with knowledge of the relative positions of the participants.

As illustrated in Figure 3, we have 8 different orientation classes $\{\omega_1, \omega_2, ..., \omega_8\}$. Let f_t be the feature vector corresponding to the data at time t as described further below, essentially a vectorized representation of the ToF depth map corresponding to a seated individual. $P(f_t|\omega_j)$ represents the likelihood of the class ω_j with respect to f_t. Section 4.3.1 describes the computation of f_t and $P(f_t|\omega_j)$ in detail. $P(\omega_j)$ represents the prior probability of the class ω_j. The method for dynamically computing the prior probability distribution at each time instant is explained in Section 4.3.2.

The posterior probabilites $P(\omega_j|f_t)$ are computed by multiplying the likelihood with the prior probabilities according to Bayes' rule:

$$P(\omega_j|f_t) = \frac{P(f_t|\omega_j)P(\omega_j)}{P(f_t)} \tag{1}$$

The class of the unknown orientation image is $j^* = \max_j\{P(\omega_j|f_t)\}$, i.e., the index of the maximum element in the vector $P(\omega_j|f_t)$.

4.3.1 The likelihood term

The 18 ToF sensors in the ceiling provide a depth map of the room. People are detected and tracked from this depth map based on computer vision techniques as described in [15]. Each tracked person blob is resized to a 10×10 image, and the distance values normalized to the range [0,1]. We then vectorize each 10×10 region into a $d = 100$ dimensional feature vector f.

Let $f_i^1, f_i^2, \dots, f_i^{n_i}$ be the feature vectors corresponding to the n_i available training orientation images corresponding to class ω_i. Our hypothesis is that the test feature vector $f_t \in R^d$ can be approximately expressed as a linear combination of the training images of the same class [27], i.e., if f_t is in class ω_i, we can express it as:

$$f_t = x_i^1 f_i^1 + x_i^2 f_i^2 + ... + x_i^{n_i} f_i^{n_i} \tag{2}$$

We construct a feature matrix, or dictionary, D using the available training images for each orientation class as:

$$D = [f_1^1, f_1^2, \dots, f_1^{n_1}, f_2^1, \dots, f_2^{n_2}, \dots, f_8^{n_8}] \tag{3}$$

In our implementation, $n_i = 100$ for all classes. Thus, $D \in R^{100\times800}$, since we have 8 classes, each with 100 training samples and each training feature vector $f_i^k \in R^{100}$. We can now express f_t in terms of all the training feature vectors of all the eight classes as:

$$f_t = Dx \tag{4}$$

where $x \in R^{800}$ is the coefficient vector. Given f_t and D, our problem is to solve the linear inverse problem in (4) to recover x.

Intuitively, if f_t corresponds to one of the classes $\{1, 2, \dots, 8\}$, the only non-zero entries in the solution vector to the above problem should be the ones that correspond to that particular class. For example, if the class of f_t is ω_i, ideally, the coefficient vector x will have the following structure: $x = [0.....0, x_i^1 ... x_i^{n_i}, 0.....0]$. The goal is to solve the above linear inverse problem so we can determine the locations of these non-zero entries in the solution vector x. Therefore, the problem of recognizing the unknown orientation image is reduced to a sparse recovery problem. Ideally, the solution can be obtained by minimizing its l_0 norm. However, since this is an NP-hard problem, we instead employ l_1 relaxation and solve the following problem:

$$\begin{aligned} &\underset{x}{\text{minimize}} && ||x||_1 \\ &\text{subject to} && f_t = Dx. \end{aligned} \tag{5}$$

The above problem belongs to a class of constrained optimization problems and can be solved by a traditional interior points method. However, this approach is slow for real-time applications like estimating the orientation of people in a room. Therefore, we convert this problem to an unconstrained basis pursuit problem using a regularization term:

$$x^* = \arg\min_x \quad ||x||_1 + \frac{1}{2\lambda}||Dx - f_t||_2^2 \tag{6}$$

This is now an unconstrained convex optimization problem that can be solved by variants of traditional gradient descent algorithms, in which the computational effort is a relatively cheap matrix-vector multiplication involving D and $D^\top$. Here, we use the Fast Iterative Shrinkage Thresholding Algorithm (FISTA) [4] to recover x^*, which is the optimum value of the coefficient vector x.

In FISTA, x is determined iteratively using the following equation:

$$x_k = \tau_\eta(y_k - \eta D^\top(Dy_k - f_t)) \tag{7}$$

where x_k is the value of x at the k^{th} iteration, $y_k = x_{k-1} + \frac{t_{k-1}-1}{t_k}(x_{k-1} - x_{k-2})$, and $\tau_\eta(x)$ is the shrinkage operator defined on each element $x(i)$ of the l-dimensional vector x as:

$$\tau_\eta(x)_i = \text{sign}(x(i))\max\{|x(i)| - \eta, 0\}, i = 1, 2, \dots, l \tag{8}$$

The factor t_{k+1} is updated as $t_{k+1} = \frac{\sqrt{1+4t_k^2}}{2}$. Thus, FISTA is similar to the Iterative Shrinkage Thresholding Algorithm (ISTA), except that it employs the shrinkage operator on the point y_k, a combination of the previous two values $\{x_{k-1}, x_{k-2}\}$.

FISTA gives the optimal coefficient vector x^* as a solution to Equation (4). After computing x^*, the likelihood of a particular class with respect to the unknown image is

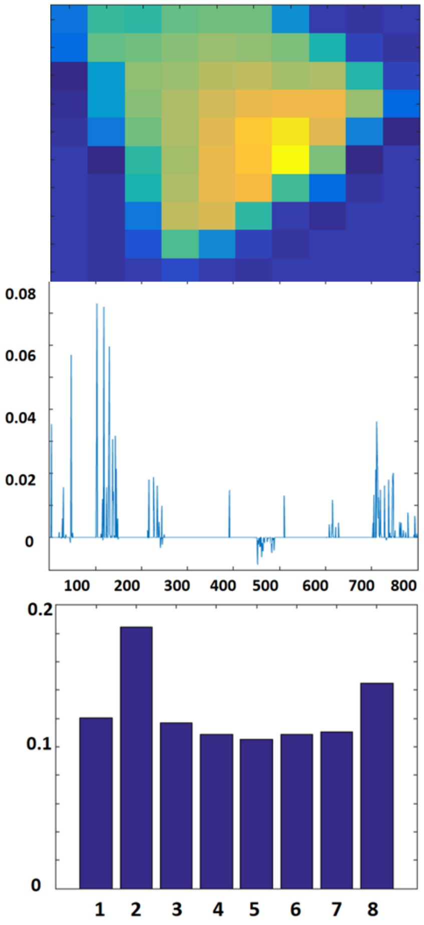

Figure 4. (a) Test orientation depth map, (b) Sparse x^* vector showing peak values corresponding to class 2, (c) The likelihood of all orientation classes with respect to the feature vector

computed from the the class-wise residuals. Let $\delta_i(x^*)$ be a vector of the same size as x^* whose only non-zero entries are the values in x^* corresponding to the class ω_i. The class-wise residuals are computed as $r_i = ||f_t - D\delta_i(x^*)||$ [27], where $i = 1, 2, \ldots, 8$. The inverses of the residuals, normalized to the range [0,1], give the likelihood $P(f_t|w_i)$ for each class.

Figure 4 illustrates an example in which Figure 4a shows the ToF depth map corresponding to a person blob. Figure 4b shows the sparse coefficient vector obtained by solving Equation (6) using FISTA, and Figure 4c shows the resulting likelihood distribution for all eight classes. From the likelihood distribution, class ω_2 is the most probable for this ToF depth map.

4.3.2 The prior term

In a typical meeting scenario, the participants are generally visually focused on the speaker. When the participants are

Algorithm 1: Computation of prior probability distribution at ToF frame t

```
Input  : ToF depth map, Speech labels, $N_p$,
         $\omega = \{1, 2, \ldots, 8\}$, $P(S)$, $P(NS)$, $P(E)$
Output: $Prior_k = \{p_1, p_2, \ldots, p_8\}$, for
         $k = 1, 2, \ldots, N_p$ at time instant t
Find weighted centroid of each person blob;
Find speaker/s $s$ from non-zero indices of speech
  labels, $N_s = |s|$;
if $N_s \neq 0$ then
    for $k \leftarrow 1, 2, \ldots, N_p$ do
        Initialize $p_i = 0, i = 1, \ldots, 8$;
        Find angle between participant $k$ and all other
          participants from their weighted centroids;
        if $k \in s$ then
            Bin each angle to nearest class label $\in \omega$,
              let this set be $\{O_k\}$;
            $p_i = p_i + 0.6/|O_k|$ , for $i \in \{O_k\}$;
            $p_i = p_i + 0.4/|\omega|$ , for $i \in \{\omega\}$;
        end
        else if $k \notin s$ then
            Bin angle/s with speaker to nearest class
              label $\in \omega$, let this set be $\{SD_k\}$;
            Bin angle/s with non-speaker/s to nearest
              class label $\in \omega$, let this set be $\{NSD_k\}$;
            Case 1: Participant looking at speaker
            $p_i = p_i + P(S) * 0.6/|SD_k|$ for
              $i \in \{SD_k\}$;
            $p_i = p_i + P(S) * 0.4/|NSD_k|$ for
              $i \in \{NSD_k\}$;
            Case 2: Participant looking at
              non-speaker
            $p_i = p_i + P(NS) * 0.7/|NSD_k|$,
              $i \in \{SD_k\}$;
            $p_i = p_i + P(NS) * 0.3/|SD_k|$ for
              $i \in \{NSD_k\}$;
            Case 3: Participant looking elsewhere
            $p_i = p_i + P(E)/|\omega|$ for $i \in \{\omega\}$;
        end
    end
end
else if $N_s = 0$ then
    for $k \leftarrow 1, 2, \cdots, N_p$ do
        Initialize $p_i = 0, i = 1, \ldots, 8$;
        Find angle between participant $k$ and all other
          participants from their weighted centroids;
        Bin each angle to nearest class label $\in \omega$, let
          this set be $\{L_k\}$;
        $p_i = p_i + 0.6/|O_k|$ , for $i \in \{L_k\}$;
        $p_i = p_i + 0.4/|\omega|$ , for $i \in \{\omega\}$;
    end
end
```

seated in swivel chairs, it is the natural tendency of the participants to orient their bodies to face the speaker. However, the amount of time participants are actually oriented towards or looking at the speaker varies from individual to individual. We used a portion of our meeting training dataset to compute the probabilities of each person looking at a speaker, looking at a non-speaker, and looking completely elsewhere. The results are tabulated in Table 1. We see that the percentage of time Person P4 is looking at a speaker is significantly less than the other participants, indicating that the actual time participants focus on the speaker is individual-specific.

The prior probability distribution for each participant is computed dynamically at each frame depending on the location of the participant with respect to the speaker(s). Algorithm 1 summarizes the steps for computing the prior probability distribution for each participant at a particular time instant. The basic idea is to compose the prior distribution such that the mass is concentrated at the orientations corresponding to the vectors from the given participant to the speaker(s), which also allows for the participant-dependent possibility of looking at a non-speaker or in a random direction. Thus, we require participant-specific values of the probabilities that the head orientation corresponds to a speaker, non-speaker, or somewhere else, denoted $P(S)$, $P(NS)$, and $P(E)$ respectively. These values are estimated from the meeting training data as discussed above.

Table 1. Measured probabilities for looking at a speaker, non-speaker or elsewhere.

	Speaker	Non-speaker	Elsewhere
P1	53.9	20.5	25.5
P2	61.8	26.1	12.1
P3	66.3	7.1	26.6
P4	33.2	24.3	42.5
Average	53.8	19.5	26.7

Figure 5 illustrates an example computation of the prior probability distribution for Person P1 when Person P4 is speaking. The algorithm first computes the relative orientation of the participants from their weighted centroids in the ToF depth map. The calculated angle is quantized into the nearest orientation class label. In Figure 5, SD_1 denotes the set of orientation directions between P1 and the speaker/s. NSD_1 denotes the set of directions between P1 and the non-speakers. Here, $SD_1 = 2$, because the angle between the centroids of P1 and P4 (the speaker) is quantized most closely to class label 2. Similarly, $NSD_1 = \{1, 3\}$ for this example.

When P1 looks at the speaker (P4), his body may not be oriented directly towards P4. We empirically estimate that the probability that P1's body is oriented towards P4 is 0.6, i.e., $P(\omega_2|S) = 0.6$. P1 can also look at P4 when his body is oriented in directions 1 or 3. Thus, we set $P(\omega_1|S) = 0.2$ and $P(\omega_3|S) = 0.2$. The other conditional probabilities are computed similarly, as illustrated in Figure 5 and Algorithm 1. As shown in Figure 5, the prior distribution is ultimately a weighted combination of three cases: P1 looks at the speaker, P1 looks at a non-speaker, and P1 looks somewhere else. Overall, the prior probability distribtion is computed as:

$$P(\omega_i) = P(\omega_i|S)P(S) + P(\omega_i|NS)P(NS) + P(\omega_i|E)P(E), i \in [1, 8] \quad (9)$$

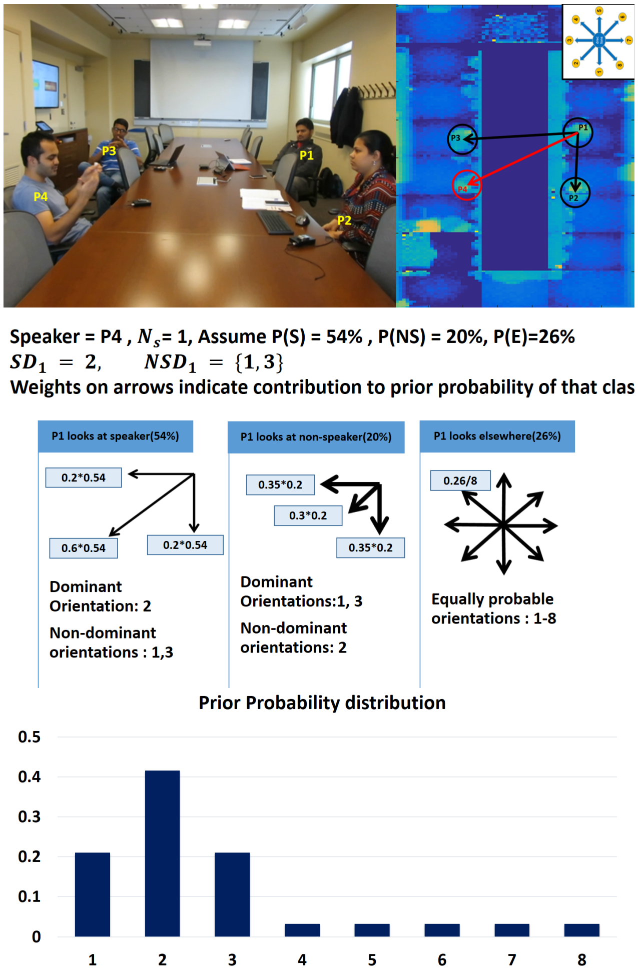
P3
P1
P4
P2
P3
P1
P4
P2
Speaker = P4 , N_s= 1, Assume P(S) = 54% , P(NS) = 20%, P(E)=26%
$SD_1 = 2$, $NSD_1 = \{1, 3\}$
Weights on arrows indicate contribution to prior probability of that clas
P1 looks at speaker(54%)
0.2*0.54
0.6*0.54
0.2*0.54
Dominant Orientation: 2
Non-dominant orientations : 1,3
P1 looks at non-speaker(20%)
0.35*0.2
0.3*0.2
0.35*0.2
Dominant Orientations:1, 3
Non-dominant orientations: 2
P1 looks elsewhere(26%)
0.26/8
Equally probable orientations : 1-8
Prior Probability distribution
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8

Figure 5. Example computation of prior probability distribution for P1 when P4 is the speaker.

From the final prior probability distribution in Figure 5, we note that the distribution has a peak at orientation direction 2 and falls off on either side of 2. Thus, the prior probability model strongly leverages information about the speaker and the relative position between the participant and the speaker. As the speaker changes from frame to frame, the prior distribution also changes. The prior distribution is also sensitive to the location of the participants and would change dynamically, e.g., if the participants shift or exchange seats in the room.

4.3.3 Median filtering

Since people are unlikely to change their seated orientations abruptly while a meeting is in progress, we median filter the estimated orientations with a filter of size 45 frames (roughly 5 seconds). This median filtering smooths out erroneous spikes in the orientation estimation.

5. Experimental Results

The meeting dataset was separated into the first 60% (approximately 7 minutes) for training and the remaining 40% (approximately 5 minutes) for testing.

The dictionary D is a 100×800 matrix whose columns are the training feature vectors. We use 100 training samples for each of the 8 classes. For each class, 80 of these 100 training samples were randomly selected from the separate non-meeting dataset described in Section 3.2. Half of these samples were collected from each of the two table positions illustrated in Figure 1. The samples collected in Table Position 2 are rotated by 180° to have the same interpretation as those in Table Position 1. To leverage actual meeting data, the remaining 20 training samples for each class were sampled randomly from the meeting training dataset. Some classes like 6, 7, and 8 (i.e., facing away from the table) are not represented in the meeting training set at all; for these classes, the dictionary was formed entirely from the non-meeting dataset. Since the selection of training samples for the dictionary D is random, the process was repeated 10 times and the accuracy of the orientation estimation algorithm on the training data set was computed in each trial. The optimized D_{opt} is the dictionary that yielded the best accuracy over all 10 trials.

We computed the accuracy of the algorithm as the percentage of the ToF frames in which the estimated orientation class is exactly equal to the actual orientation class. We also calculated the percentage of frames in which the actual class and the estimated class differ by 1 bin, i.e., the difference between the exact and estimated orientation is less than 45°. As we noted during our manual annotation of body orientations, even human observers can disagree to ±1 class.

Figure 6 illustrates a sample result of the orientation classification algorithm. Figure 6a is the reference view, Figure 6b is the corresponding raw ToF data, and Figure 6c shows the location, speaker, and estimated body orientations. A short video clip with the reference camera view, the ToF raw output, and the algorithm results are available at the link `https://youtu.be/Hm98ZEqjAtk`. Figure 7 shows the estimated and the actual body orientations of person P1 for the entire meeting duration, showing excellent correspondence.

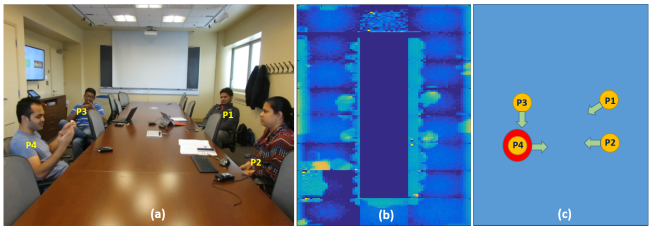

Figure 6. (a) Camera view, (b) Raw data from the ToF sensors stitched to form a depth map of the room, (c) Results of the orientation estimation algorithm: the red circle indicates the speaker, detected from the microphone recordings. The yellow arrows indicate the automatically estimated body orientations.

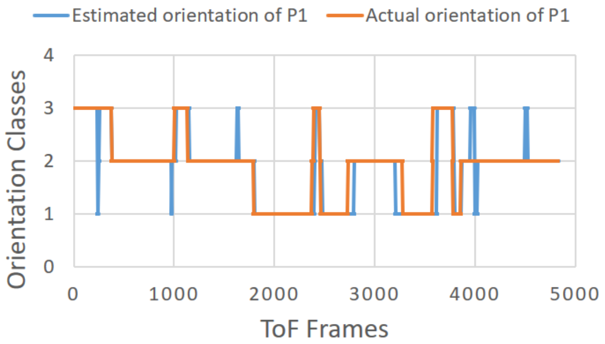

Figure 7. Actual and estimated body orientations of person P1 for the entire meeting duration.

Table 2 tabulates the orientation classification accuracy on the training and testing datasets respectively. The accuracy on the training dataset is 90.5% and on the testing dataset is 80.8%. The percentage of cases where the actual and estimated orientation differ by 1 class is 6% on the training dataset and 17.2% on the testing dataset. Therefore, the algorithm correctly predicts the orientation of seated individuals to an accuracy of ±45° for 96% of the total time for the training dataset and 98% for the test data set.

Before finalizing the orientation classification algorithm, we ran several experiments on our training data set, which can be interpreted as different versions of our algorithm. The accuracy increased in each version and the algorithm presented in Section 4 is our final and best version. Figure 8 shows the accuracy of the different versions of the algorithm on the training data set. The versions are listed below:

Version 1: D was formed entirely from randomly selected samples of the non-meeting training dataset. No prior location/speaker information was added, thus making all 8 classes equally probable. The classification algorithm is essentially reduced to maximum likelihood estimation.

Table 2. Orientation estimation accuracy with dictionary D_{opt} and location and speaker based priors.

Training Data Set						Testing Data Set				
	P1	P2	P3	P4	Average	P1	P2	P3	P4	Average
Differ by 0 class	93.3	94.1	91.2	83.4	90.5	88	65.3	89.2	80.7	80.8
Differ by at most 1 class	99.5	99.2	93	94.3	**96.5**	97.7	99.4	95	100	**98**

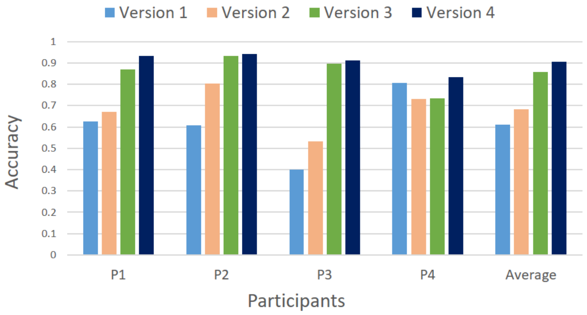

Figure 8. The accuracy of different versions of the orientation estimation algorithm.

Version 2: The same D as in Version 1 was used. The prior probabilities were modeled based on relative locations of the participants without any speaker/audio information.
Version 3: The same D as in Version 1 was used. The prior probabilities were modeled based on location and speaker information as described in Section 4.3.2.
Version 3: The formation of the matrix D was modified to leverage actual meeting data as described in the beginning of this section. The optimal dictionary D_{opt} was used. The prior probabilities were based on location and speaker information, as in Version 3.

From Figure 8, we note that the average classification accuracy increases by about 7% when introducing priors based on location alone and by about 17% when modeling prior probabilities based on both location and speaker information. Introducing training samples from the actual meeting to form the dictionary gives an additional 4% boost to the average accuracy. Thus, the final version of the algorithm gives an improvement of approximately 30% over the first version. We also noted that the median filtering at the end of the algorithm improved the average accuracy by approximately 3–7%, as compared to the same algorithm without any filtering.

6. Conclusions and Future Work

We described a method for estimating the seated orientation of individuals using a fusion of time-of-flight and audio data. The system can be used for group meeting analysis and facilitation, in which a smart room needs to know the exact location, pose, and orientation of each participant.

At present, we are using individual lapel microphones for recording the audio information. In the future, we intend to replace these microphones with a custom 16-channel ambisonic (spherical) microphone [8, 9]. 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, and make the sensing even less obtrusive.

Another constraint in this dataset is that the participants' positions were basically fixed for the entire duration of the meeting. In a more natural meeting scenario, a participant may walk up to the board to present something, leave the meeting early, or get up and sit in a different seat. We plan to integrate these realistic scenarios in our future experiments and test the robustness of our algorithm.

Finally, we want to integrate the location, pose, and orientation information with other verbal and non-verbal cues to detect the visual focus of attention of the group, determine interaction links between participants, and study productivity and participation shifts in a group meeting. We believe such smart rooms that provide accurate time-stamped information of participants' location, pose, orientation, and speech would be of immense value to social psychologists who study group dynamics in real physical environments.

7. Acknowledgement

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. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We also extend our thanks to Tianna-Kaye Woodstock and Li Jia.

References

[1] S. Afshari, T. Woodstock, M. Imam, S. Mishra, A. Sanderson, and R. Radke. The Smart Conference Room: An Integrated System Testbed for Efficient, Occupancy-Aware Lighting Control. In *ACM Int. Conf. Embedded Syst. Energy-Efficient Built Environments*, Seoul, S.Korea, 2015.

[2] Audacity. Audacity. `http://www.audacityteam.org/`, 2017. [Online; accessed 09-March-2017].

[3] S. O. Ba and J.-M. Odobez. Multiperson Visual Focus of Attention from Head Pose and Meeting Contextual Cues. *IEEE Trans. Pattern Anal. and Machine Intelligence*, 33(1):101–116, 2011.

[4] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[5] C. Beyan, N. Carissimi, F. Capozzi, S. Vascon, M. Bustreo, A. Pierro, C. Becchio, and V. Murino. Detecting Emergent Leader in a Meeting Environment Using Nonverbal Visual Features Only. In *Proc. ACM Int.Conf. Multimodal Interaction*, Tokyo, Japan, 2016.

[6] S. Burger, V. MacLaren, and H. Yu. The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style. In *INTERSPEECH*, Denver, CO, 2002.

[7] N. Campbell, T. Sadanobu, M. Imura, N. Iwahashi, S. Noriko, and D. Douxchamps. A Multimedia Database of Meetings and Informal Interactions for Tracking Participant Involvement and Discourse Flow. In *Proc. LREC*, Genoa, Italy, 2006.

[8] S. Clapp, A. Guthrie, J. Braasch, and N. Xiang. Three-Dimensional Spatial Analysis of Concert and Recital Halls with a Spherical Microphone Array. In *ASA Proc. Meetings Acoust.*, Montreal, Canada, 2013.

[9] S. Clapp, A. E. Guthrie, J. Braasch, and N. Xiang. Headphone-and Loudspeaker-based Concert Hall Auralizations and Their Effects on Listeners' Judgments. *The J. of the Acoust. Soc. of America*, 134(5):3969–3969, 2013.

[10] D. Gatica-Perez. Automatic Nonverbal Analysis of Social Interaction in Small Groups: A Review. *Image and Vision Computing*, 27(12):1775–1787, 2009.

[11] D. Gatica-Perez, A. Vinciarelli, and J.-M. Odobez. Nonverbal Behavior Analysis. In *Multimodal Interactive Syst. Manage.*, pages 165–187. EPFL Press, 2014.

[12] T. Giannakopoulos and A. Pikrakis. *Introduction to Audio Analysis: A MATLAB® Approach*. Academic Press, 2014.

[13] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, et al. The ICSI Meeting Corpus. In *IEEE Int. Conf. Acoust., Speech, and Signal Process.*, Hong Kong, China, 2003.

[14] D. Jayagopi, D. Sanchez-Cortes, K. Otsuka, J. Yamato, and D. Gatica-Perez. Linking Speaking and Looking Behavior Patterns with Group Composition, Perception, and Performance. In *Proc. ACM Int. Conf. Multimodal Interaction*, Santa Monica, CA, 2012.

[15] L. Jia and R. J. Radke. Using Time-of-Flight Measurements for Privacy-Preserving Tracking in a Smart Room. *IEEE Trans. Ind. Informat.*, 10(1):689–696, 2014.

[16] N. Jovanović, A. Nijholt, et al. Addressee Identification in Face-to-Face Meetings. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[17] N. Jovanovic, R. op den Akker, and A. Nijholt. A Corpus for Studying Addressing Behaviour in Multi-Party Dialogues. *Language Resources and Evaluation*, 40(1):5–23, 2006.

[18] S. Mathur, M. S. Poole, F. Pena-Mora, M. Hasegawa-Johnson, and N. Contractor. Detecting Interaction Links in a Collaborating Group Using Manually Annotated Data. *Social Networks*, 34(4):515–526, 2012.

[19] J. F. Nunamaker Jr, R. O. Briggs, D. D. Mittleman, D. R. Vogel, and B. A. Pierre. Lessons From a Dozen Years of Group Support Systems Research: A Discussion of Lab and Field Findings. *J. Manage. Inform. Syst.*, 13(3):163–207, 1996.

[20] C. Oertel, K. A. Funes Mora, S. Sheikhi, J.-M. Odobez, and J. Gustafson. Who Will Get the Grant?: A Multimodal Corpus for the Analysis of Conversational Behaviours in Group Interviews. In *Proc. ACM Workshop Understanding Modeling Multiparty, Multimodal Interactions*, Istanbul, Turley, 2014.

[21] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato. A Realtime Multimodal System for Analyzing Group Meetings by Combining Face Pose Tracking and Speaker Diarization. In *Proc. ACM Int. Conf. Multimodal Interfaces*, Crete, Greece, 2008.

[22] K. Otsuka, H. Sawada, and J. Yamato. 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. ACM Int. Conf. Multimodal Interfaces*, Aichi, Japan, 2007.

[23] K. Otsuka, Y. Takemae, and J. Yamato. A Probabilistic Inference of Multiparty-Conversation Structure Based on Markov-Switching Models of Gaze Patterns, Head Directions, and Utterances. In *Proc. ACM Int. Conf. Multimodal Interfaces*, Trento, Italy, 2005.

[24] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase. Conversation Scene Analysis with Dynamic Bayesian Network Based On Visual Head Tracking. In *Proc. IEEE Int. Conf. Multimedia and Expo*, Toronto, ON, Canada, 2006.

[25] D. Sanchez-Cortes, O. Aran, and D. Gatica-Perez. An Audio Visual Corpus for Emergent Leader Analysis. In *Workshop Multimodal Corpora Mach. Learning: Taking Stock and Road Mapping the Future*, Alicante, Spain, 2011.

[26] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling Focus of Attention for Meeting Indexing Based on Multiple Cues. *IEEE Trans. Neural Networks*, 13(4):928–938, 2002.

[27] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust Face Recognition Via Sparse Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* , 31(2):210–227, 2009.