Parallel Detection of Conversational Groups of Free-Standing People and Tracking of their Lower-Body Orientation

Marynel Vázquez[1,2], Aaron Steinfeld[1] and Scott E. Hudson[2,3]

***Abstract*— Appropriate robot behavior in public, open spaces cannot occur without the ability to automatically detect conversational groups of free-standing people. To this end, we propose an alternating optimization procedure that estimates lower body orientations and detects groups of interacting people. The first task is achieved by tracking the direction of the lower body of the people in the scene based on their position, their head orientation, the location of objects of interest in their vicinity, and their groups. For the second task, we propose a new group detection algorithm based on F-formation detection. This method can reason about lower body orientation distributions, and generates soft group assignments for the orientation trackers. We evaluate the proposed approach on a publicly available dataset, and show that it can improve state-of-the-art detection of non-interacting people without sacrificing group detection accuracy. This is particularly useful for robots since it provides more opportunities for starting interactions and can help estimate disengagement.**

I. INTRODUCTION

We want to improve robot reasoning of human environments and social behaviors by giving them the ability to detect conversational groups of free-standing people from visual data. For example, in Figure 1(a) two people converse and the robot is not part of the interaction. Thus, it would be inappropriate for the robot to pass between them or suddenly approach very closely. Meanwhile, the robot has the opportunity to interact with the other people in the room. In Figure 1(b), the robot interacts with four people. If the robot can detect that one person is leaving the group but wants to continue engaging this person, then it can resort to strategies to solicit his or her attention. Thus, detecting group membership can help robots better conform to social norms, as well as reason about spatial relationships in public environments [1], [2]. In addition, detecting these type of social encounters can enable autonomous attention shifts [3] and initiate human-robot interactions [4], [5].

Existing algorithms for detecting conversational groups exploit the fact that the members tend to cooperate to sustain a shared focus of attention, and maintain a particular spatial-orientational organization that maximizes their opportunities to monitor each other's mutual perceptions [6], [7]. The methods that reason directly about foci of attention typically seek to detect the intersection of the lines of sight of the people in the scene [8]–[10]. While this information can be used to infer group interactions, inferences are often affected by shifts in attention, temporary visual distractions (e.g., when people glance at someone who walks nearby) or situations in which people do not necessarily converse, but pay attention to the same event or target. Other methods that reason about the spatial-orientational organizations of free-standing group interactions [5], [11]–[15] tend to be more robust. The reason is that these organizations persist over time, even when the members of a group change or people attend to distractions. It is also possible to detect group interactions using proximity only [16], [17], since people stand nearby while conversing [18]. However, distance information alone can lead to errors in constrained spaces (e.g., in crowds). Besides, [19] proposed a data-driven approach to detect dyadic group interactions. Unfortunately, detecting bigger groups with this method requires to learn new models.

Fig. 1: We seek to detect free-standing conversational groups in situations such as (a) and (b). In (a), two people (outlined in white) sustain a face-to-face arrangement while talking. In (b), the group interacts with a furniture robot. The transactional segments are illustrated in blue on the floor. The o-space is where the transactional segments intersect.

Kendon [20] described the spatial-orientational organizations that emerge during free-standing conversations as face-formation systems, or **F-Formations** in short. The members of such systems position and orient themselves such that they have *equal, direct*, and *exclusive* access to the space between them. In the case of pairs, the individuals typically sustain face-to-face or side-by-side arrangements during F-Formations (as in Fig. 1(a)). Bigger groups tend to form semi-circular, square, or circular arrangements (Fig. 1(b)).

The methods that reason about F-formations to detect conversational groups typically model the *transactional segment* of each person in the scene and then reason about *o-spaces*. The transactional segment of a person is the space that extends forward from his or her **lower body** and is used while engaged in a particular activity [20]. The extent of this space varies, but is always limited by the orientation of their lower body, as in Fig. 1. People often look into this space, and actively maintain it in the presence of intrusions.

Group members tend to orient themselves so their individual transactional segments intersect, creating a joint transactional space. This intersection is known as the o-space of the corresponding F-formation, and is maintained even when the interactants get distracted briefly and turn

[1] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. {marynel,steinfeld}@cmu.edu

[2] Disney Research Pittsburgh, Pittsburgh, PA 15213, USA.

[3] HCI Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. scott.hudson@cs.cmu.edu

their heads away. When people join an existing group, they typically position themselves such that their transactional segments intersect the existing o-space. As in [2]–[5], the notion of F-formations can be applied to robots as well.

While prior F-formation detection methods were designed based on the previous concepts, most of them used head orientations or an estimate of focus of attention to direct people's transactional segments [11], [12], [14], [15]. The reason is twofold: head orientations and focus of attention approximate lower body orientation, and are easier to estimate automatically.[1] Unfortunately, this choice makes the F-formation detection methods prone to the aforementioned problems, requiring additional effort to handle noise in the direction of the transactional segments. As a trade-off, F-formation detection algorithms have to become more inclusive in order to succeed, resulting in additional false members in the detected groups (as further discussed in our evaluation). Alternatively, one might consider relying on the direction of motion of the individuals in the scene to model their transactional segments, since this has worked for tracking moving groups [21]. Nonetheless, this information is insufficient by itself when people stand still while conversing.

Building from previous work, we propose to detect groups using an estimate of the (non-observable) lower body orientations of the people in the scene, rather than the head directly. We exploit the mutual dependency between the group detection task and the lower body orientation estimation problem to this end. On one hand, lower body orientations allow us to estimate F-formation systems and, thus, infer who is conversing with whom. On the other, o-spaces are an important prior for inferring lower body orientations. Our algorithm processes each problem in parallel as part of an alternating optimization procedure that quickly approximates the lower body orientation distribution of everybody in the scene, while estimating their corresponding groups and o-spaces. This makes our approach easy to implement (without any additional image processing) and suitable for in situ human-robot interaction applications, where computational resources are often limited and on-line inference is preferred.

As part of our contributions, we also provide a new method for detecting F-formations based on o-spaces. This method can work with non-parametric orientation distributions and is able to compute soft group assignment scores for each person in the scene, rather than the most likely group configuration only. The latter property is particularly useful for reasoning about group interactions in confusing situations.

We test our approach on a publicly available dataset [14], in which six people hold various conversations for about 30 min. By inferring lower-body orientations in this scenario, we can better detect when people are not part of a nearby group conversation in comparison to the state-of-the-art method of Setti et al. [15]. This can increase the number of opportunities for a robot to start new interactions with people, without sacrificing group detection accuracy.

[1] In the case of [5], a laser scanner was used to measure users' upper-body orientation with respect to a robot and detect triadic F-formations. This approach required direct view of the upper bodies from the sensor, limiting generalization to other situations. In [13], body skeletons from multiple Kinects were used to detect F-formations, but orientation measurements needed manual correction when frontal and backward skeletons were mislabeled. This situation could be improved by tracking orientations based on body measurements and contextual data, as proposed in this work.

II. NOTATION

In this work, we focus on estimating a set of conversational groups $\mathcal{G}_t = \{(G_1, \mathbf{c}_1), (G_2, \mathbf{c}_2), \ldots\}$ at any time t by means of detecting F-formations. We model each group with a set G that holds the numeric identifiers of its members and a 2D vector $\mathbf{c} = [c_x, c_y]^T$ with the location of its o-space center in a world coordinate frame. For example, if the first group has three members, then $G_1 = \{a, b, c\}$, where $a, b, c \in \mathbb{N}$ are the identifiers of the interactants.

At any time t, we assume that we are given a set $\mathcal{O}_t = \{\mathbf{o}_t^1, \ldots, \mathbf{o}_t^O\}$ with the 2D location of the objects that people may engage with in their vicinity (e.g., as taken from a map of the environment). We are also given the 2D locations $\mathcal{P}_t = \{\mathbf{p}_t^1, \ldots, \mathbf{p}_t^P\}$ of the people in the scene and their head orientations $\Theta_t = \{\theta_t^1, \ldots, \theta_t^P\}$ (yaw angle) with respect to the world frame. These values come from an arbitrary person tracker and, as such, are prone to measurement errors.

We often use von Mises distributions ($\mathcal{VM}$) in this work for the lower body orientation estimation problem because they naturally model angular distributions [22]. In particular, $\mathcal{VM}(a; \mu, \kappa) = \exp(\kappa \cos(a - \mu))/2\pi I_0(\kappa)$, with $I_0(\cdot)$ the modified *Bessel* function of order zero. The parameters μ and κ are analogous to the mean and the inverse of the variance in the normal distribution. When $\kappa = 0$, the von Mises distribution becomes uniform.

III. DETECTING GROUPS AND TRACKING LOWER-BODY ORIENTATIONS

Our algorithm GRUPO, for GROUP detection and Orientation tracking, alternates between each task (Fig. 2). The algorithm outputs at each time-step a set of free-standing conversational groups G_t by detecting F-formations, as well as lower-body orientation distributions for the people in the scene. Information about the likely o-space centers is output by the F-formation detection module and fed into the lower-body orientation trackers (one for each person) along with their corresponding positions $\mathbf{p}$, head orientations θ, and the location of nearby objects of interest. Each tracker then outputs a lower-body orientation distribution that is used to guide the group detection process in the next time-step.

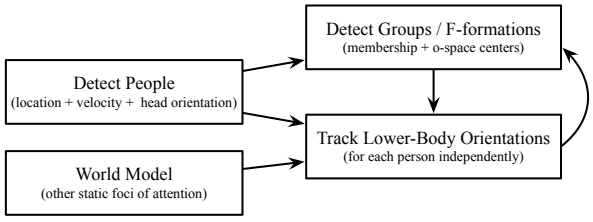


Fig. 2: GRUPO alternates between detecting free-standing conversational groups and tracking lower-body orientations.

GRUPO requires good initialization, as any other local optimization method. We use head orientations to compute a first set of groups and likely o-space centers, and then pass this information to the orientation trackers. The next two sections detail the main steps of the algorithm.

IV. DETECTING FREE-STANDING CONVERSATIONAL GROUPS

We propose a new method for detecting F-formations and their members, inspired by the Hough voting scheme of [12].

The main difference with prior work is that our method reasons about non-parametric lower body orientation distributions. In other words, our method can cope with situations of high uncertainty, where the distributions are multi-modal. In addition, the proposed F-formation detection approach computes soft o-space assignments in a continuous space. These are good properties because soft assignments help lower-body orientation trackers recover from group detection errors, and operating in a continuous space frees our method from quantization problems. This makes our method faster and more accurate than earlier voting approaches, like [12].

Our F-formation detection approach is detailed in Algorithm 1. First, each person i proposes o-space centers (modeled as normal distributions) based on his or her position $\mathbf{p}^i$ in the scene and lower-body orientation distribution Φ^i (provided as a set of N samples). The proposals compose a Gaussian mixture whose 2D local maxima $\mathcal{M}$ we estimate using the fixed-point mode-finding algorithm of [23] (lines 9 to 25 of Alg. 1). Considering the modes of the Gaussian mixture as likely o-space centers, we then compute a set of scores $\mathcal{S}^i[k]$ for every person i in the scene. These scores represent the likelihood that the transactional segment of this person intersects the o-space center k (line 41). When it is likely that the transactional segment of a person intersects at least one o-space center, these scores compose a discrete probability distribution that represents soft o-space assignments. Otherwise, all $\mathcal{S}^i[k] = 0$ for any person i by convention. Finally, we group the people for whom a particular mode has the highest probability and return these groupings as the detected set of F-formations (with associated o-space centers).

A. PROPOSED O-SPACE CENTERS

Without loss of generality, consider a person i with a corresponding lower-body orientation distribution $\Phi^i = [\phi^i[1], \ldots, \phi^i[N]]$, with each $\phi^i[\cdot] \in [0, 2\pi]$ with respect to the horizontal axis of the world coordinate frame. We model the proposed o-space center for a given $\phi^i[j]$, $1 \leq j \leq N$, as a 2D normal distribution $\mathcal{N}(\mu^i_j, \Sigma^i_j)$ with:

$$\mu^i_j = \mathbf{p}^i + R \begin{bmatrix} stride + f(abs(\mathbf{d1}^T \mathbf{v}^i)) \\ 0 \end{bmatrix} \quad (1)$$

$$\Sigma^i_j = R \begin{bmatrix} \sigma^j_x & 0 \\ 0 & \sigma^j_y \end{bmatrix} R^T \quad (2)$$

$$\text{where, } R = \begin{bmatrix} \cos(\phi^i[j]) & -\sin(\phi^i[j]) \\ \sin(\phi^i[j]) & \cos(\phi^i[j]) \end{bmatrix}$$

$$\mathbf{d1} = \begin{bmatrix} cos(\phi^i[j]) \\ sin(\phi^i[j]) \end{bmatrix}$$

The vectors $\mathbf{p}^i$ and $\mathbf{v}^i$ are the position and velocity of the person i with respect to the world coordinate frame, and the $stride$ parameter represents the likely distance of the o-space center the lower body when the person is standing still. The function $f : \mathbb{R}^+ \rightarrow [0, d]$ is monotonically increasing, and we use it to increment the distance between the person and the proposed o-space center up to an arbitrary maximum value $d \in \mathbb{R}^+$, based on the individual's velocity and the direction $\mathbf{d1}$ of the lower body. Our rationale for this model is that people often move forward when approaching an existing group. When they are already interacting, they sometimes

Algorithm 1: Detect F-formations by mode-finding

Input: Position $\mathbf{p}^i$ and non-parametric lower-body orientation distribution $\Phi^i = [\phi^i[1], \ldots, \phi^i[N]]$ of every person i in the scene ($1 \leq i \leq P$)
Output: Groups $\mathcal{G}$, list $\mathcal{M}$ of possible o-space centers, and lists $\mathcal{S}^i$ of o-space scores for every person

1 $\mathcal{X} = \emptyset$ // set of mixture components
2 $w = 1/PN$ // components' weight
3 **for** $i = 1$ *to* P **do**
4 **for** $j = 1$ *to* N **do**
5 $(\mu^i_j, \Sigma^i_j) = ospaceProposal(\mathbf{p}^i, \phi^i[j])$
6 $\mathcal{X} = \mathcal{X} \cup \{(\mu^i_j, \Sigma^i_j, w)\}$
7 **end**
8 **end**
9 $\mathcal{M} = []$ // modes (possible o-spaces)
10 **for** $(\mu^i_j, \Sigma^i_j, w^i_j)$ *in* $\mathcal{X}$ **do**
  // hill climb from the mean [23]
11 $\mathbf{x} = fixedPointLoop(\mu^i_j, \mathcal{X})$
12 **if** $\mathbf{x}$ *is local maxima* **then**
13 $(idx, dist) = closestMode(\mathbf{x}, \mathcal{M})$
14 **if** $dist < \tau$ **then** // group modes?
15 **if** $p(\mathcal{M}[idx]; \mathcal{X}) < p(\mathbf{x}; \mathcal{X})$ **then**
  // x has higher probability
16 $\mathcal{M}[idx] = \mathbf{x}$
17 **end**
18 $k = idx$
19 **else** // add new mode
20 add $\mathbf{x}$ to $\mathcal{M}$
21 $k = |\mathcal{M}|$
22 **end**
23 $mode\_idx^i_j = k$ // bookkeeping
24 **end**
25 **end**
  // compute soft assignment scores
26 **for** $i = 1$ *to* P **do**
27 $\mathcal{S}^i = []$
28 **for** $k = 1$ *to* $|\mathcal{M}|$ **do** // initialization
29 $n^i_k = 0$
30 add 0 to $\mathcal{S}^i$
31 **end**
32 **for** $j = 1$ *to* N **do**
33 **if** $isset(mode\_idx^i_j)$ **then**
  // reached local maxima
34 $k = mode\_idx^i_j$
35 **if** $visible(\mathcal{M}[k], \mathbf{p}^i)$ **then**
36 $n^i_k = n^i_k + 1$
37 **end**
38 **end**
39 **end**
40 **if** $\sum_k n^i_k > 0$ **then**
41 **for** $k = 1$ to $|\mathcal{M}|$ **do** $\mathcal{S}^i[k] = n^i_k / \sum_k n^i_k$ **end**
42 **end**
43 **end**
  // greedy hard group assignment
44 $\mathcal{G} = \emptyset$
45 **for** $k = 1$ *to* $|\mathcal{M}|$ **do**
46 $G = \emptyset$
47 **for** $i = 1$ *to* P **do**
  // get the most-likely o-space
48 $idx = \arg\max_m \mathcal{S}^i[m]$
49 **if** $\mathcal{S}^i[idx] > 0$ *and* $k == idx$ **then**
50 $G = G \cup \{i\}$
51 **end**
52 **end**
53 **if** $|G| \geq 2$ **then** // found group / F-formation
54 $\mathcal{G} = \mathcal{G} \cup \{(G, \mathcal{M}[k])\}$
55 **end**
56 **end**

move sideways or backward to allow other people to join their F-formation, without altering much their o-space.

In terms of the function f in eq. (1), we use $f(x) = 2\sigma(x) - 1$ with $\sigma(x) = 1/(1 + \exp(-x))$, though other functions could be used as well. With our choice, the o-space can move maximum 1m away from a person that is moving forward or backwards. The o-space moves little when the person walks sideways, since $abs(\mathbf{d1}^T\mathbf{v}^i)$ approaches zero.

We further control the shape of Σ_j^i in eq. (2) with:

$$\sigma_x^j = (stride/s)^2 + g(abs(\mathbf{d1}^T\mathbf{v}^i)) \text{ and } \sigma_y^j = \lambda(stride/s)^2 \tag{3}$$

with s, $\lambda \in \mathbb{R}^+ - \{0\}$, and g another increasing function. Figure 3 illustrates the flexibility of this model.

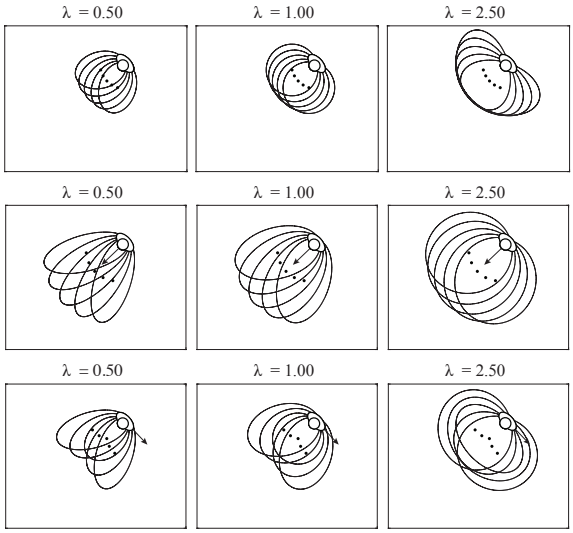


Fig. 3: O-space proposals for 5 orientations at $0, \pm 0.25, \pm 0.5$ radians from the direction of the lower body. The distributions were computed with $stride = 0.7$m, $s = 3$, $g(x) = f(0.5x)$ and various λ (eq. (1)-(3)). The velocity was zero for the first row, it was aligned with the direction of the lower body in the second, and was perpendicular to it in the third. The black dots are the means of the Gaussian distributions. Ellipses represent the covariances at 99% confidence.

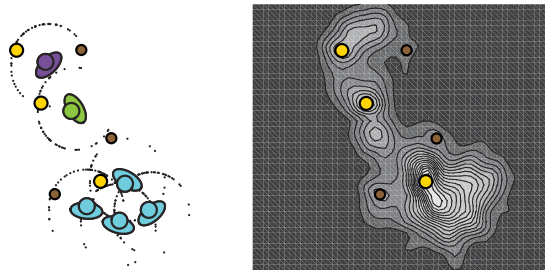
B. O-SPACE MIXTURE DISTRIBUTION

O-space proposals are combined into a Gaussian mixture:

$$p(\mathbf{x}) = \sum_i^P \sum_j^N \frac{1}{NP} \mathcal{N}(\mathbf{x}; \mu_j^i, \Sigma_j^i) \tag{4}$$

where $1/NP$ is the weight of the components, and μ_j^i and Σ_j^i come from equations (1) and (2), respectively.

We consider as possible o-space centers the modes of this mixture (Fig. 4). To find the modes, we use the fixed-point algorithm of [23], starting from the means of the components. The function $fixedPointLoop$ in line 11 of Alg. 1 corresponds to the "fixed point iteration loop" of [23] (see their Fig. 3 for more details). As in the latter work, we decide in line 12 whether a sample point $\mathbf{x}$ reached a local maxima based on the Hessian of the mixture at that point. While there can be more modes than mixture components and it is possible that a more exhaustive search is required to find them all, sampling from the means provides good results in practice with a reduced computational load.



(a) Spatial configuration (b) Contour plot

Fig. 4: *Left:* shows the means of each person's o-space proposals (small black dots) and groups (by color) based on the corresponding mode probability. *Right:* Mixture distribution of o-space proposals. The modes (in yellow and brown) were found for $\tau = 0.75$m (line 14 of Alg. 1). Yellow modes were the most likely o-space for at least one person.

We group the modes that are within τ meters from each other (line 14 of Alg. 1), and keep track of which component converged to which mode in the process. When this grouping happens, we set the mode with highest mixture probability as the most-likely o-space center in its vicinity. In this manner, the parameter τ helps coping with noise in human motion, as well as in our estimates of lower-body orientation.

Once the likely o-space centers are found, we count for each person how many of their mixture components converged per center, and compute their o-space scores by normalizing this count (line 41 of Alg. 1). In order to ensure that the members of an F-formation have direct access to the o-space, we do not consider in the count of each person the o-space centers that are not directly visible from his or her position. For computing visibility, we model people as circumferences with fixed radius (0.2m) and compute occlusions by ray-casting. The resulting soft group assignment scores are passed to the orientation tracker of the corresponding person.

To obtain hard group assignments, we proceed in a greedy fashion and pick the mode with highest score as the most likely o-space center per person. A group is set to be found whenever a possible o-space center has the highest score for two or more people (line 53 of Alg. 1).

V. TRACKING LOWER-BODY ORIENTATIONS

We pose the estimation of lower-body orientation as a tracking problem, based on the following observations: (1) people tend to orient their lower body towards other people or objects of interest while standing still, (2) people often orient their head in the same direction as their lower body, (3) people can turn their heads (temporarily) to attend to visible targets other than their main focus of attention, and (4) people tend to orient their lower body towards their direction of motion while walking. In general, we assume that people are standing at all times, as it happens during free-standing conversations, and that the likely o-space centers and corresponding assignment scores for each person are given (e.g., as output by our group detection algorithm).

A. RECURSIVE STATE ESTIMATION

At time t, we estimate (independently) the probability distribution of each person i's lower body orientation ϕ_t^i using the dynamic Bayesian Network of Fig. 5. We assume that the person's velocity $\mathbf{v}^i$, position $\mathbf{p}^i$, head orientation θ^i, and contextual information C^i are given for this estimation process (from time step 1 up to t). The contextual information includes the position $\mathbf{p}^j$ (where $j \neq i$) of the other people in the scene, the set $\mathcal{O}$ with the locations of the nearby objects that people my interact with, the o-space centers $\mathcal{M}$ and the assignment scores $\mathcal{S}^i[k]$, for $1 \leq k \leq |\mathcal{M}|$.

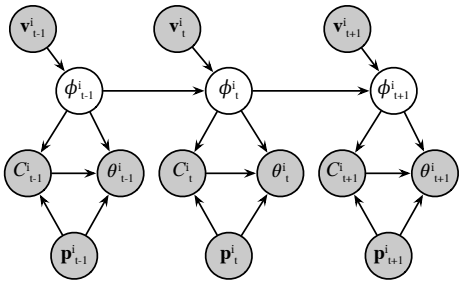


Fig. 5: Bayes network that characterizes the evolution of hidden lower body orientation ϕ^i of a person i, based on his or her position $\mathbf{p}^i$, linear velocity $\mathbf{v}^i$, head orientation measurement θ^i, and contextual information C^i.

The belief $bel(\phi_t^i)$ at time t can be formulated recursively:

$$bel(\phi_t) = p(\phi_t|\mathbf{v}_{1:t}, \theta_{1:t}, C_{1:t}, \mathbf{p}_{1:t}) = \tag{5}$$
$$\eta\, p(\theta_t|\phi_t, C_t, \mathbf{p}_t) p(C_t|\phi_t, \mathbf{p}_t) \int p(\phi_t|\phi_{t-1}, \mathbf{v}_t) bel(\phi_{t-1}) d\phi_{t-1}$$

where we have dropped momentarily the superscript i for simplicity. In this factorization, η is a normalization term, $p(\theta_t|\phi_t, C_t, \mathbf{p}_t)$ is the *head measurement probability*, $p(C_t|\phi_t, \mathbf{p}_t)$ is the *context probability*, and $p(\phi_t|\phi_{t-1}, v_t)$ is the *state transition probability*.

We use a particle filter to approximate the posterior $bel(\phi_t)$ with a finite number of samples $\Phi_t = \big[\phi_t[1], \ldots, \phi_t[N]\big]$, which we initialize from a uniform $\mathcal{VM}(0, 0)$. At any following time step t, we follow Algorithm 2 to update the belief. First, we sample new particles given our transition probability and our previous distribution Φ_{t-1} (line 3 of Alg. 2). Then, we compute a weight w_t, or importance factor, for each particle based on our context and head measurement probabilities (line 4). Finally, we draw particles with replacement based on the weights (lines 7-10). We use low variance sampling in practice for this last step [24].

Motion Model: For any person i, we propagate his or her lower-body orientation ϕ^i from time $t-1$ to t as follows:

$$\phi_t^i = \phi_{t-1}^i + \omega(\mathbf{v}_t^i, \phi_{t-1}^i)\Delta T + q \tag{6}$$

The angular velocity function $\omega(\mathbf{v}_t, \phi_{t-1})$ in eq. (6) controls the rate of rotation of the lower body, ΔT is the time difference from $t-1$ to t, and $q \sim \mathcal{N}(0, r)$ is a small perturbation. The angular velocity function changes based on the person's motion and orientation:

$$\omega(\mathbf{v}_t^i, \phi_{t-1}^i) = \text{sign}(\mathbf{d2}^T \mathbf{d3}) \Big[\frac{\alpha}{\Delta T}\Big] m(\mathbf{v}_t^i, \alpha) \tag{7}$$

Algorithm 2: Particle filter for lower-body orientation

Input: $\Phi_{t-1}, \mathbf{v}_t, C_t, \theta_t$
Output: Φ_t
1 $\bar{\Phi}_t = \Phi_t = [\,]$
2 **for** $j = 1$ *to* N **do**
3 sample $\phi_t[j] \sim p(\phi_t|\phi_{t-1}[j], \mathbf{v}_t)$
4 $w_t[j] = p(\theta_t \mid \phi_t[j], C_t, \mathbf{p}_t)\, p(C_t \mid \phi_t[j], \mathbf{p}_t)$
5 add $(\phi_t[j], w_t[j])$ to $\bar{\Phi}_t$
6 **end**
7 **for** $j = 1$ *to* N **do**
8 draw k with probability $\propto w_t[j]$
9 add $\phi_t[k]$ from $\bar{\Phi}_t$ to Φ_t
10 **end**

where,

$$\begin{aligned}
\alpha &= \arccos(\mathbf{d1}^T \mathbf{d3}) \\
\mathbf{d1} &= [\cos(\phi_{t-1}^i)\ \ \sin(\phi_{t-1}^i)]^T \\
\mathbf{d2} &= [-\sin(\phi_{t-1}^i)\ \ \cos(\phi_{t-1}^i)]^T \\
\mathbf{d3} &= \mathbf{v}_t^i / \|\mathbf{v}_t^i\| \\
m(\mathbf{v}_t^i, \alpha) &= 2\sigma(h(\alpha)\|\mathbf{v}_t^i\|) - 1
\end{aligned} \tag{8}$$

The variable α is the (unsigned) angular difference between the previous lower-body orientation ϕ_{t-1}^i and the current direction of motion on the ground plane. The $\text{sign}(\mathbf{d2}^T \mathbf{d3})$ component of (7) provides the direction of rotation of the lower body as the person walks. The geometric relations between $\mathbf{d1}$, $\mathbf{d2}$ and $\mathbf{d3}$ are illustrated in Figure 6.

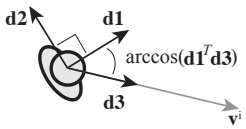


Fig. 6: Geometric relations in our motion model.

The function $m(\mathbf{v}_t^i, \alpha)$ in eq. (8) is used to scale the angular velocity. Our particular choice returns a number in $[0, 1)$ for non-negative inputs, using the sigmoid function $\sigma(\cdot)$ as in eq. (1). The function $h(\alpha)$ in eq. (8) returns a non-negative scaling constant that further controls the effect of the magnitude of the linear velocity on ω, based on the direction of motion. In practice, we use $h(\alpha) = a\frac{\pi-\alpha}{\pi} + b$, with $a > b > 0$ small fixed constants. This means that when the person moves backwards ($\alpha \to \pi$), the function h returns a small value and, thus, m is also small. In this manner, the motion model favors small changes in body orientation, in comparison to sudden rotations of 180°.

Context Model: The probability of the context C_t^i at time t given the lower body orientation ϕ_t^i and position $\mathbf{p}_t^i$ is a mixture of three probabilities:

$$p(C_t^i|\phi_t^i, \mathbf{p}_t^i) = \begin{bmatrix} w_{\text{group}} \\ w_{\text{eng}} \\ (1 - (w_{\text{group}} + w_{\text{eng}})) \end{bmatrix}^T \begin{bmatrix} p_{\text{group}}(C_t^i|\phi_t^i, \mathbf{p}_t^i) \\ p_{\text{eng}}(C_t^i|\phi_t^i, \mathbf{p}_t^i) \\ \mathcal{VM}(0; \phi_t^i, 0) \end{bmatrix}$$

where the weights are positive or zero, and sum to one. The component $p_{\text{group}}(C_t^i|\phi_t^i, \mathbf{p}_t^i)$ is the probability of the o-space assignment given the person's spatial configuration, $p_{\text{eng}}(C_t^i|\phi_t^i, \mathbf{p}_t^i)$ is the probability of engagement with another person or object, and $\mathcal{VM}(0; \phi_t^i, 0)$ is a uniform distribution that represents our failure to explain the context.

The probability $p_{\text{group}}(C_t^i|\phi_t^i, \mathbf{p}_t^i)$ is another mixture based on the o-space centers $\mathcal{M}_t$ and the scores $\mathcal{S}_t^i$:

$$p_{\text{group}}(C_t^i|\phi_t^i, \mathbf{p}_t^i) = \sum_{k=1}^{|\mathcal{M}_t|} \mathcal{S}_t^i[k]\mathcal{VM}(\beta_k; \phi_t^i, \kappa_{\text{group}}) + \Big(1 - \sum_{k=1}^{|\mathcal{M}|} \mathcal{S}_t^i[k]\Big)\mathcal{VM}(0; \phi_t^i, 0) \quad (9)$$

where β_k is the angle of the unitary vector $(\mathcal{M}[k] - \mathbf{p}_t^i)/\|\mathcal{M}[k] - \mathbf{p}_t^i\|$, and $\kappa_{\text{group}} > 0$ is a parameter that controls the spread of the von Mises distributions. The last term of eq. (9) is very important in two cases: when $\mathcal{M}$ is empty; and when the person's transactional segment is not intersecting any known o-space center (the values of $\mathcal{S}_t^i$ are all zero).

We model $p_{\text{eng}}(C_t^i|\phi_t^i, \mathbf{p}_t^i)$ similarly:

$$p_{\text{eng}}(C_t^i|\phi_t^i, \mathbf{p}_t^i) = \sum_{v=1}^{V} e_v \, \mathcal{VM}(\beta_v; \phi_t^i, \kappa_{\text{eng}}) \quad (10)$$

where β_v is the direction from person i to the other people and objects of interest that are directly visible within a field of view of 180°. The weights e_v in (10) satisfy $\sum_{v=1}^{V} e_v = 1$, and we use them to bias $p_{\text{eng}}(c_t|\phi_t)$ based on Hall's spatial zones [18]:

$$e_v = dist\_weight(d_v) / \sum_{q=1}^{V} dist\_weight(d_q) \quad (11)$$

where with d_v the distance (in meters) to the person or object v in C_t^i. The function $dist\_weight(d_v)$ returns 0.6 if the distance is in the personal or intimate spaces ($d_v <= 1.2$), 0.3 if the distance is in the social space ($1.2 < d_v <= 3.6$), 0.1 if the distance is in the public space ($3.6 < d_v < 7.6$), and 0 otherwise. In the exceptional case that C_t^i contains no visible person or object of interest within the upper range of the public space, we evaluate $p_{\text{eng}}(C_t^i|\phi_t^i, \mathbf{p}_t^i)$ with a uniform distribution (as we do for p_{group} when $\mathcal{M}$ is empty).

Head Measurement Model: The probability is given by:

$$p(\theta_t^i|\phi_t^i, C_t^i, \mathbf{p}_t^i) = w_{\text{front}}\, p_{\text{front}}(\theta_t^i|\phi_t^i) + w_{\text{focus}}\, p_{\text{focus}}(\theta_t^i|\phi_t^i, C_t^i, \mathbf{p}_t^i) + (1 - w_{\text{front}} - w_{\text{focus}})\, \mathcal{VM}(\theta_t; 0, 0)$$

where the weights normalize the mixture once again. The first component p_{front} accounts for frontal headings, the second describes the head orientation based on possible foci of attention (or distractions), and the third accounts for unexplained head orientation measurements. In particular,

$$p_{\text{front}}(\theta_t^i|\phi_t^i) = \mathcal{VM}(\theta_t^i; \phi_t^i, \kappa_{\text{front}})$$

with κ_{front} the spread of the distribution. The probability

$$p_{\text{focus}}(\theta_t^i|\phi_t^i, C_t^i, \mathbf{p}_t^i) \propto \max_{v=1\ldots V} \{\mathcal{VM}(\theta_t^i; \beta_v, \kappa_{\text{focus}})\}$$

is proportional to the maximum likelihood of orienting the head towards a (non-occluded) person, object of interest, or most likely o-space center within a 180° field of view in front of person i. We set p_{focus} to $\mathcal{VM}(\theta_t^i; 0, 0)$ if no focus of attention is visible.

Note that people may interact with other social entities that can be added to this model, such as other robots. It is also possible to incorporate information about how certain we are about the location of these targets through the κ parameters, though we use constant values in this work.

VI. EVALUATION

We compare the performance of the proposed algorithms against the state-of-the-art approach of [15]. We used their implementation[2] to generate the results in this paper.

Dataset: We used the Cocktail Party dataset of [14] for our evaluation. This dataset consists of a sequence of more than 24000 images (recorded at 15Hz) that show six people interact in an instrumented room. Besides the images, the dataset provides the location of each person and their head orientation as computed by a custom person tracker, as well as ground truth group annotations for 320 frames (roughly every 5 seconds).

We collected annotations for the lower body orientation of the people in the scene to complement the dataset. The annotations were made on the 320 images that had group annotations, using an interface similar to the one that was used in [25] to collect body orientations.[3]

Group detection criteria: We adopt the two criteria of [12], [14], [15] for analyzing group detection results versus the ground truth annotations. In one case, we consider a group to be detected if at least $\lceil (2/3)|G| \rceil$ of its members are identified and no more than $1 - \lceil (2/3)|G| \rceil$ of false subjects are found, where $|G|$ is the cardinality of the group. In a harder case, we consider a group to be detected if all its members are identified and no false members are found. Precision, recall and F1 scores are computed using this criteria, summing true positives, false positives, and false negatives over all the frames with group annotations.

Parameterization: As in prior work, we used $stride = 0.7$m for our group detection algorithm and the one from [15]. We considered a table that was in the scene of the cocktail party as an object that people could interact with (modeled by two landmarks). In the case of Alg. 1, we also used $\lambda = 0.25$ and $g(x) = f(0.5x)$ in eq. (3), which we found to work well in practice.

We ran the particle filters for GRUPO with $N = 80$ samples, and set the variance $r = 0.2$, and the parameters $a = 0.64$ and $b = 0.16$ for sampling the motion model. The weights of the context and head measurement probabilities were $w_{\text{group}} = 0.55$, $w_{\text{eng}} = 0.35$, $w_{\text{front}} = 0.2$ and $w_{\text{focus}} = 0.75$, respectively. Finally, $\kappa_{\text{group}} = 2$, $\kappa_{\text{eng}} = 4$, $\kappa_{\text{front}} = 3$ and $\kappa_{\text{focus}} = 5$, which provided good results experimentally.

To generate results for the proposed group detection method (Alg. 1) using head measurements and lower body annotations, we generated a (non-parametric) distribution Φ of lower body orientations by sampling $\mathcal{N}(\phi, q)$, with ϕ the head or body angle and q a small variance. With this approach, $N = 30$ samples sufficed to obtain good results, with $q = 0.07$ and $q = 0.13$ for the lower body orientations and head measurements, respectively. As expected, a higher variance worked better for the head measurements in comparison to the lower body orientations.

A. DETECTING GROUPS WITH GRAPH CUTS [15]

We first analyzed the effect of the maximum description length (MDL) parameter on the performance of [15]. We ran

[2] http://profs.sci.univr.it/~cristanm/ssp/

[3] The annotations can be downloaded from http://cs.cmu.edu/~marynelv/cocktailparty_lborient.txt.

this group detection method using head measurements, lower body annotations and the estimated lower body direction with GRUPO. For the latter case, we replaced Alg. 1 in GRUPO with [15]. The soft o-space scores used by our particle filters were set to binary $\{0, 1\}$ values depending on the detected groups (since [15] only provides hard group assignments).

Figure 7a shows F1 scores. The higher the MDL parameter, the more penalty is given to the number of detected groups (see eq. (5) in [15]). There is not a unique MDL parameter that works best for all input orientations. In fact, the more noise, the higher the MDL should be. But there is a trade-off: the higher MDL, the more inclusive the algorithm becomes, grouping people together more often than not.

Table I shows the best results obtained with [15]. Using lower body annotations results in better performance, followed by using the head directly and GRUPO's estimate. Our intuition as to why GRUPO does not improve the results is that [15] only computes the most likely group assignment, and often includes false members in the detected groups when the input orientations are noisy. Each of these properties can lead to errors that propagate within GRUPO.

TABLE I: Best group detection results using [15]. "LB Ann" is lower body annotations. Results for GRUPO were averaged over 5 runs (std. errors were equal to or less than 0.005).

Criteria	Orientation	MDL	Precision	Recall	F1
$\lceil(2/3)\|G\|\rceil$	LB Ann	14000	0.84	0.84	0.84
	Head	30000	0.82	0.81	0.82
	GRUPO	14000	0.82	0.80	0.81
$\|G\|$	LB Ann	14000	0.69	0.68	0.69
	Head	30000	0.62	0.61	0.61
	GRUPO	14000	0.61	0.60	0.61

B. DETECTING GROUPS WITH ALGORITHM 1

We then evaluated the performance of the proposed group detection method (Alg. 1). Figure 7b shows typical F1 scores when the parameter s varies, which we use to control σ_x in equation (3). The smaller s, the more spread the o-space proposal distributions along the direction of the body. In terms of F1 scores, Algorithm 1 is as powerful as the graph cuts approach of [15] when lower body annotations are used for orienting the transactional segments. The results for the head are slightly lower in this case, but GRUPO tends to perform better under the full group detection criteria. Table II provides the precision, recall and F1 scores for the best parameter s in each case. Figure 8 shows illustrative results.

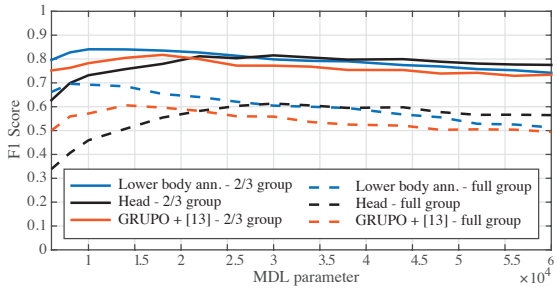
TABLE II: Best group detection results using Alg. 1. "LB Ann" is lower body annotations. Results were averaged over 5 runs (std. errors were equal to or less than 0.003).

Criteria	Orientation	s Param	Precision	Recall	F1
$\lceil(2/3)\|G\|\rceil$	LB Ann	2	0.86	0.83	0.85
	Head	1.25	0.81	0.80	0.81
	GRUPO	2	0.82	0.80	0.81
$\|G\|$	LB Ann	2	0.71	0.69	0.70
	Head	1.25	0.60	0.59	0.60
	GRUPO	2	0.65	0.63	0.64

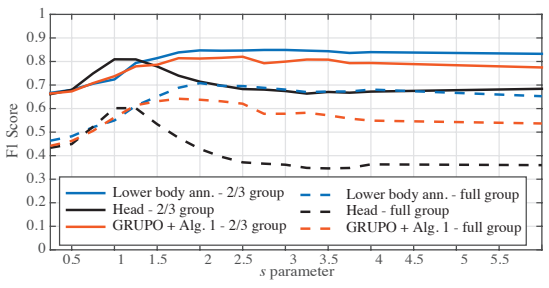
Fig. 8: Qualitative results for GRUPO on the Cocktail Party dataset. The first column shows ground truth groups (by color) and lower body orientations. The second one shows group detections using the head ($s = 1.25$ in eq. (3)). The third uses estimated lower body orientations ($s = 2$).

C. ORIENTATION ESTIMATION

Figure 9 shows a histogram of the absolute angular difference between lower body orientation annotations and head measurements, and between the annotations and the estimated lower body directions on a typical run of GRUPO. We used Alg. 1 and $s = 2$ (as in Tab. II) to compute these results. On average, the head measurements were 0.59 radians ($\sim 34°$) off from the body annotations (SE=0.013).



(a) Graph cuts group detection [15]



(b) Proposed group detection (Alg. 1)

Fig. 7: *Left:* Group detection results for [15], using various maximum description length (MDL) parameters. *Right:* Results for the proposed group detection method, for various values of s (eq. (3)). Continuous lines correspond to detecting at least $\lceil(2/3)|G|\rceil$ group members; dashed lines are for detecting all members (and no more).

Using GRUPO, the estimated lower body orientations were 0.38 radians ($\sim$ 22°) on average from the annotations (SE=0.008). GRUPO tends to better approximate real lower body orientations than raw head orientation measurements.

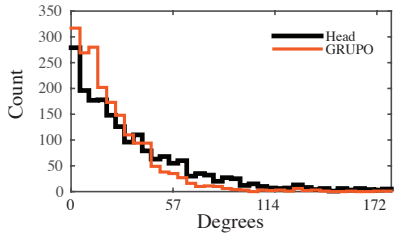


Fig. 9: Superimposed histograms of the absolute angular difference from the head and GRUPO's orientation to the lower body annotations. Bins have a width of 5°.

D. INDIVIDUAL INTERACTION DETECTION

Finally, we looked at how well [15] could infer if people were interacting or not versus GRUPO. Table III shows the results from this binary classification task, where accuracy is $(TP + TN)/(TP + FP + TN + FN)$ and the true negative rate is $TN/(TN + FP)$, with TP the number of true positives, TN the true negatives, FP the false positives, and FN the false negatives. While GRUPO and the method of [15] have similar accuracy at the individual level, GRUPO is able to double the true negative rate of [15], without any additional computer vision processing.

TABLE III: Individual interaction classification results. We used MDL= 30000 for [15], and $s = 2$ (eq. (3)) for GRUPO. The latter results were averaged over 5 runs.

Metric	GC (Head) [15]	GRUPO
True Positives	1739	1707.4 (SE = 2.0)
False Positives	140	97.4 (SE = 1.1)
True Negatives	39	81.6 (SE = 1.1)
False Negatives	2	33.6 (SE = 2.0)
Accuracy	0.93	0.93 (SE < 0.01)
True Neg. Rate	0.22	0.46 (SE < 0.01)

VII. CONCLUSIONS

The proposed method detects conversational groups by recognizing F-formations and tracking lower body orientations. The latter is advantageous for improving how the transactional segments are modeled, even if body orientations are not directly observed. Our results suggest that this approach can help better detect non-interacting people, without sacrificing group detection performance. We expect this method to produce better engagement estimates by robots and more opportunities for them to start social interactions.

There are additional opportunities to improve GRUPO by providing it additional information, such as direct measurements of lower body orientations. In this case, we foresee the particle filters reasoning about head and body orientations jointly, as in [26]. Furthermore, temporal constraints could be added to the F-formation detections, and GRUPO could estimate groups and track orientations at different frequencies. This would help deal with reduced computational resources.

We plan on testing GRUPO on a more complex dataset [27] and will pursue additional algorithm enhancements in the future. We believe that integrating lower-body orientation tracking and group detection will lead to improved robot autonomy and more appropriate social behavior.

ACKNOWLEDGMENTS

We thank The Walt Disney Corporation for their support of this research effort. We also thank O. Lanz for the Cocktail Party dataset, M. Cristani for their group detection code [15], and E. J. Carter for her assistance on this project.

REFERENCES

[1] P. Marshall, Y. Rogers, and N. Pantidi, "Using F-formations to Analyse Spatial Patterns of Interaction in Physical Environments," in *Proc. CSCW*, 2011.
[2] H. Huettenrauch, K. Severinson Eklundh, A. Green, and E. Topp, "Investigating spatial relationships in human-robot interaction," in *Proc. IROS*, 2006.
[3] H. Kuzuoka, Y. Suzuki, J. Yamashita, and K. Yamazaki, "Reconfiguring Spatial Formation Arrangement by Robot Body Orientation," in *Proc. HRI*, 2010.
[4] C. Shi, M. Shimada, T. Kanda, H. Ishiguro, and N. Hagita, "Spatial formation model for initiating conversation," in *Proc. RSS*, 2011.
[5] M. A. Yousuf, "Mobile Museum Guide Robots Able to Create Spatial Formations with Multiple Visitors," Ph.D. dissertation, Saitama University, Saitama, Japan, 9 2013.
[6] E. Goffman, *Behavior in public places: Notes on the social organization of gatherings*. Free Press of Glencoe, 1963.
[7] E. Goffman, P. Drew, and A. Wootton, *Erving Goffman: Exploring the Interaction Order*. Polity Press, 1988.
[8] A. Fathi, J. Hodgins, and J. Rehg, "Social interactions: A first-person perspective," in *Proc. CVPR*, 2012.
[9] H. S. Park, E. Jain, and Y. Sheikh, "3D Social Saliency from Head-Mounted Cameras," in *Proc. NIPS*, 2012.
[10] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino, "Social Interactions by Visual Focus of Attention in a Three-Dimensional Environment," *Expert Systems*, vol. 30, no. 2, pp. 115–127, 2013.
[11] H. Hung and B. Kröse, "Detecting F-formations As Dominant Sets," in *Proc. ICMI*, 2011.
[12] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino, "Social Interaction Discovery by Statistical Analysis of F-Formations," in *Proc. BMVC*, 2011.
[13] T. Gan, Y. Wong, D. Zhang, and M. S. Kankanhalli, "Temporal Encoded F-formation System for Social Interaction Detection," in *Proc. of MM*, 2013.
[14] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani, "Multi-Scale F-Formation Discovery for Group Detection," in *Proc. ICIP*, 2013.
[15] F. Setti, C. Russell, C. Bassetti, and M. Cristani, "F-formation Detection: Individuating Free-standing Conversational Groups in Images," *CoRR*, vol. abs/1409.2702, 2014.
[16] T. Yu, S.-N. Lim, K. Patwardhan, and N. Krahnstoever, "Monitoring, recognizing and discovering social networks," in *Proc. CVPR*, 2009.
[17] M. E. J. Newman, "Modularity and community structure in networks," *Proc. of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.
[18] E. T. Hall, *The Hidden Dimension*. Anchor Books, 1990.
[19] R. Mead and M. J. Matarić, "Automated Proxemic Feature Extraction and Behavior Recognition: Applications in Human-Robot Interaction," *Int'l Journal of Social Robotics*, vol. 5, no. 3, pp. 367–378, 2013.
[20] A. Kendon, *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge Univ. Press, 1990.
[21] M. Luber and K. O. Arras, "Multi-Hypothesis Social Grouping and Tracking for Mobile Robots." in *Proc. RSS*, 2013.
[22] N. Fisher, *Statistical Analysis of Circular Data*. Cambridge University Press, 1995.
[23] M. A. Carreira-Perpinan, "Mode-finding for mixtures of Gaussian distributions," Dept. of Computer Science, University of Sheffield, UK, Tech. Rep. CS-99-03, 1999.
[24] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*, ser. Intelligent robotics and autonomous agents. MIT Press, 2005.
[25] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *Proc. CVPR*, 2011.
[26] F. Flohr, M. Dumitru-Guzu, J. Kooij, and D. Gavrila, "A Probabilistic Framework for Joint Pedestrian Head and Body Orientation Estimation," *IEEE Trans. Intell. Trans. Sys.*, vol. PP, no. 99, pp. 1–11, 2015.
[27] M. Vázquez, E. J. Carter, J. A. Vaz, J. Forlizzi, A. Steinfeld, and S. E. Hudson, "Social group interactions in a role-playing game," in *Proc. HRI Extended Abstracts*, 2015.