MGpi: A Computational Model of Multiagent Group Perception and Interaction

Navyata Sanghvi
Carnegie Mellon University
Pittsburgh, Pennsylvania
navyatasanghvi@cmu.edu

Ryo Yonetani*
Carnegie Mellon University
Pittsburgh, Pennsylvania
ryo.yonetani@sinicx.com

Kris Kitani
Carnegie Mellon University
Pittsburgh, Pennsylvania
kkitani@cs.cmu.edu

ABSTRACT

Toward enabling next-generation robots capable of socially intelligent interaction with humans, we present a **computational model** of interactions in a social environment of multiple agents and multiple groups. The Multiagent Group Perception and Interaction (MGpi) network is a deep neural network that predicts the appropriate social action to execute in a group conversation (*e.g.*, speak, listen, respond, leave), taking into account neighbors' observable features (*e.g.,* location of people, gaze orientation, distraction, *etc.*). A central component of MGpi is the Kinesic-Proxemic-Message (KPM) gate, that performs social signal gating to extract important information from a group conversation. In particular, KPM gate filters incoming social cues from nearby agents by observing their body gestures (kinesics) and spatial behavior (proxemics). The MGpi network and its KPM gate are learned via imitation learning, using demonstrations from our designed **social interaction simulator**. Further, we demonstrate the efficacy of the KPM gate as a social attention mechanism, achieving state-of-the-art performance on the task of **group identification** without using explicit group annotations, layout assumptions, or manually chosen parameters.

KEYWORDS

Social agent models; Socially interactive agents; Agent-based analysis of human interactions; Social group identification; Multiagent learning; Learning from demonstrations.

ACM Reference Format:
Navyata Sanghvi, Ryo Yonetani, and Kris Kitani. 2020. MGpi: A Computational Model of Multiagent Group Perception and Interaction. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 10 pages.

1 INTRODUCTION

In order to develop next-generation robots that can interact socially with humans, there is first a need for expressive computational models that can encode social intelligence [2, 19, 29, 59]. Broadly, an agent's social intelligence is its ability to understand and respond appropriately to others, *i.e.*, its: (1) social perception and (2) social interaction management skill. Social perception is the ability to analyze other agents' social signals including non-verbal behavioral information such as facial or postural expressions [41, 47], physical distance (*proxemics*) [21] and body gestures (*kinesics*) [4]. Social interaction management is the ability to take appropriate action in response to information retrieved via social perception.

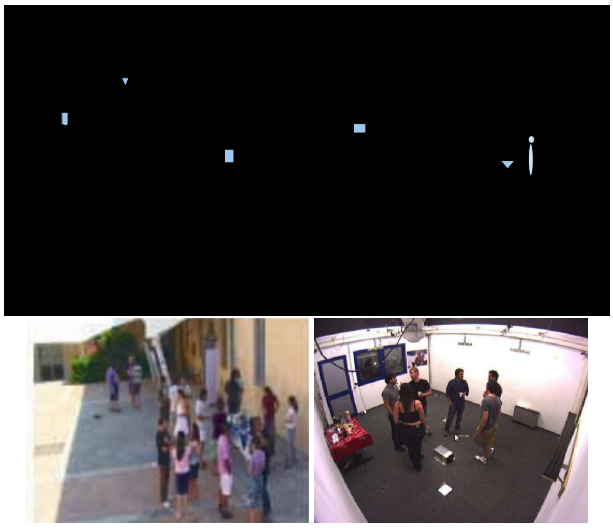

Figure 1: Top: Multiagent multigroup social intelligence at a party. Bottom: Coffee Break [15] and Cocktail Party [62] dataset images.

Toward the ultimate goal of robotic social intelligence, we propose a computational model of social interaction management based on social perception of behavioral signals in the context of multiagent multigroup conversation. Consider a crowded social party like the ones depicted in Figure 1 to understand some important components of social intelligence.

(1) **Group Formation and Social Signal Gating.** When many people are gathered together, it is natural for people to form smaller groups and engage in conversation. In order to participate effectively in a group conversation, each person needs to be able to pay attention to the people in the group, while suppressing information (both verbal and non-verbal) from other nearby groups. This type of social attention mechanism, which we term social signal gating, is a critical aspect of social perception and is closely tied to group identification (discussed later).
(2) **Dynamic Group Size and Influence.** At a social party, people dynamically move from group to group to start new conversations and leave old ones. The ability to adapt to dynamic group size and a dynamic degree of influence from nearby people is an important aspect of social intelligence.

*Ryo Yonetani is currently at OMRON SINIC X, Tokyo, Japan.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

(3) **Short-Term Memory.** Communication within a group is successful when there is an appropriate balance of conversational actions. At every moment, each group member decides their future conversational actions based on the past course of interaction (*e.g.*, speak, respond or leave). For example, if one person has been talking very long, group members might become disinterested and disengage from the group. In other words, a social agent's short-term memory of past non-verbal and conversational interactions with neighbors is influential in determining its future social interactions.

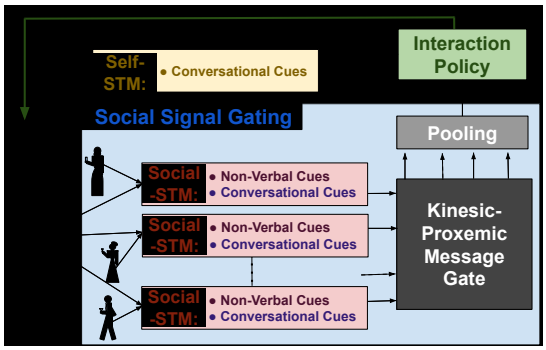

Figure 2: Overview of MGPI Network

The example above highlights critical features of a multiagent multigroup interaction model. Keeping in mind these features, we propose **MGPI**, a **M**ultiagent **G**roup **P**erception and **I**nteraction Network. As shown in Figure 2, MGPI is a deep neural network consisting of the following modules:

(1) **Social Signal Gating Module** expresses appropriate social perception by processing incoming and past social signals from neighboring agents (both non-verbal and conversational). It decides their degree of influence using a social attention function, the Kinesic-Proxemic-Message (KPM) gate. To enable reasoning with a dynamic number of influencing neighbors, it uses a pooling operation to aggregate encoded social signals.

(2) **Short-term Memory (STM) Encoders** encode the agent's short-term memory of past interactions with each neighbor (Social-STM) and its own actions (Self-STM).

(3) **Interaction Policy Module** processes outputs from aforementioned components to decide the agent's next conversational action.

The MGPI network enables us to model social agents that make action choices in a decentralized manner. The parameters of MGPI are learned end-to-end via imitation learning, using a set of demonstrated group interaction sequences. In this work, demonstrations are generated by a social interaction simulator, but MGPI can also be learned from annotated real-world social interactions. Since modeling the full extent of social intelligence is very complex (*e.g.*, power, dialogue, trust), in this work, we use abstracted expressions of social iteraction (*e.g.*, discrete conversational actions). We believe this simplification is a necessary step towards designing more complex models of real-world social interactions.

As mentioned earlier, social perception is closely tied to group identification. Thus, we further hypothesize that, if our proposed social signal gating module is modeled correctly, we will be able to map the learned social attention directly to group membership. We demonstrate the ability of the KPM gate to identify groups in a direct, unsupervised manner, achieving state-of-the-art performance against group detection methods that use hand-defined parameters, features and complex spatial assumptions on how people arrange themselves in social situations (F-formations [34]).

In summary, our contributions are as follows:

(1) We present MGPI, a **computational model** of multiagent multigroup social interactions (Section 3).

(2) In a dearth of prior work on multiagent multigroup social interaction simulators and in the absence of real-world datasets with multigroup behavioral annotations (*e.g.*, listen, respond, strongly address), we design our own **social interaction simulator**, seeking inspiration from several works which study the multi-modal nature of small-group conversational dynamics [27, 28, 30, 33, 40]. We use this simulated data to train and evaluate our network (Section 5).

(3) We demonstrate how the ability of **group identification** *emerges* as a result of learning a social interaction policy, achieving competitive performance against state-of-the-art methods with the *explicit* aim of detecting groups. Unlike prior work, we do so without the use of explicit group annotations, layout assumptions or hand-defined parameters (Sections 4, 6).

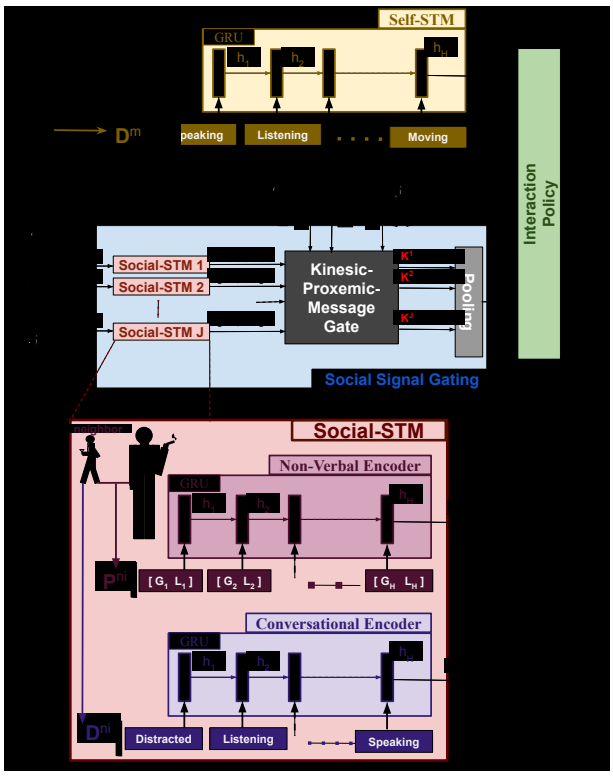

Figure 3: Components of MGPI Network: (1) Social Signal Gating Module, (2) Self-STM Encoder, (3) Interaction Policy. A Social-STM Encoder within the Social Signal Gating Module is shown in detail.

2 RELATED WORK

The analysis of intelligent agent grouping, behavior and communication is of broad interest to many disciplines including Human-Robot Interaction [6, 7, 10, 23, 46, 47, 61], Multiagent Systems and Machine Learning [5, 9, 24, 25, 36, 38, 43, 44, 48, 54, 55, 60], Neurobiology [8, 13, 18, 26, 31, 52], Computer Vision [3, 16, 17, 50, 51, 57, 62] and Psychology [2, 4, 20, 21, 41, 42]. Since this work is focused on developing computational models of multiagent communication and identifying social groups, we elaborate on selected related work.

Human-Robot Interaction: The effects of gaze and proxemics in human-robot interaction and social signal perception has been widely explored [23]. Recognizing these non-verbal cues' social importance, we incorporate the perception of such cues as a central component in our model. Dialogue systems work [49, 53] concentrates on appropriately responding to conversational cues, while our work additionally deals with the non-verbal aspects of human communication. Detection and automatic analysis of a human's engagement in social human-robot scenarios has been comprehensively studied [10, 46, 47, 61] using various non-verbal cues such as body attitudes, facial video signals, and personality traits. Related to this, our work presents a method to automatically infer an interacting agent's attention from non-verbal and conversational cues. Sociable robots have been defined and studied in the context of their degree of interaction with humans and their functionality [6, 19]. Toward enabling socially intelligent robots, we present a model of multiagent multigroup interaction. The importance of simulation theory and imitative interaction in the social understanding of neighbors by robots has been studied [7]. We use this insight in order to learn a computational model for socially intelligent robots through imitation of simulated interactions.

Multiagent Systems and Machine Learning: Emergent collective behaviors have been studied in multiagent organized swarms [5]. By contrast, we model individual behavior in a socially intelligent multigroup setting. Adversarial deceptive interactions have been studied between individuals [60] and groups of intelligent robots [48]. Instead, we assume a cooperative scenario, where interactions imitate multigroup human-to-human social reactions. Modeling multiagent communication is an active area of research in machine learning [9, 54]. Recent work makes use of deep reinforcement learning or imitation learning to discover policies for agent-to-agent communication, *e.g.*, [24, 25, 36, 38, 43, 44, 55]. The motivation for these works is often collaboratively solving a particular task which might involve partial observability. Communication protocols best fit to accomplish that task are learned. By contrast, we learn to imitate social communication protocols of humans in a multigroup scenario. Some works assume the communication of internal states between agents, while others learn to communicate a particular 'message' in order to solve the task. Neighbor messages might be aggregated using a pre-defined pooling operation. By contrast, we explicitly encode particular social signals between agents and learn signal gating to appropriately weight incoming signals, showing superior performance to prior pooling strategies.

Neurobiology and Signal Gating: The concept of social signal gating introduced in our model is inspired by the neurobiological process of *sensory gating* [18, 26]. Sensory gating is the ability to filter out unnecessary or irrelevant external stimuli. The cocktail party effect [8] is one example of auditory sensory gating where a person is able so focus auditory attention to a specific target while ignoring other irrelevant audio input. Similar sensory gating is observed in other senses as well, to prevent overwhelming the primary cortical areas [13, 31, 52]. Recognizing the importance of modelling sensory gating in socially intelligent computational systems, we use our KPM gate to mimic this phenomenon when considering the importance of nearby agents' social signals.

Social Group Interactions: For the purpose of evaluating our model, there are no large-scale public datasets of states and conversational actions of people interacting in multiple groups. Existing datasets [15, 45, 62] only provide sequences of physical positions, gaze directions and group assignments. More importantly, no conversational action annotations (*e.g.*, listening, speaking, *etc*) are provided. Closely related to our ideas, a social simulator PsychSim [39] adopts a theory-of-mind approach to modelling interactions in scenarios with agent attributes such as power and hardship. By contrast, we consider scenarios of face-to-face interactions in social situations, and require agents to have conversational roles such as responding to being addressed. Further, we aim to automatically learn the levels of reasoning required to behave appropriately in a social situation. In the absence of prior work on simulators or real-world datasets with conversational action annotations, we design our own social interaction simulator (Section 5). Inspired by several works which study the multi-modal nature of small-group conversational dynamics [27, 28, 30, 33, 40], we design simulated interaction rules in order to generate data to train our network.

Social Group Identification: There is much work on identifying the members of a group in a social situation [32, 50, 51, 58], and more recently [56]. This prior work typically assumes agents' arrangement in spatial layouts called F-formations [34], often attempting to find *o-space* centers [34] using heuristic strategies and hand-crafted parameters. By contrast, under no such layout assumptions, we learn parameters for a straightforward, automatic group identification mechanism (Section 6).

3 SOCIAL INTERACTION MANAGEMENT: THE MGPI NETWORK

Our goal is to design a computational model for multiagent multigroup interactions that incorporates signal gating, adaptation to dynamic group sizes and short-term memory encoding. Towards this goal, we propose the MGPI architecture (Figures 2, 3). In our design of MGPI, we consider several insights from social science literature: the importance of 'situational awareness' and 'presence' [2], and the impact of non-verbal social cues 'kinesics' [4] and 'proxemics' [21].

Notation. We denote the m-th agent by A^m and its J_t neighbors at time t by $\{A^{n_i}\}_{i=1}^{J_t}$. At time t, A^m and its neighbor A^{n_i} have gaze directions $\boldsymbol{g}_t^m, \boldsymbol{g}_t^{n_i} \in \mathbb{R}^2$, and positions $\boldsymbol{l}_t^m, \boldsymbol{l}_t^{n_i} \in \mathbb{R}^2$. $R(\phi)$ is the rotation matrix associated with angle $\phi = \arctan(\boldsymbol{g}_t^m)$. Relative gaze direction and relative rotated position of A^{n_i} w.r.t. A^m are respectively given by:

$$\boldsymbol{g}_t^{(n_i \leftarrow m)} = R(\phi)\boldsymbol{g}_t^{n_i} \qquad \text{and} \qquad \boldsymbol{l}_t^{(n_i \leftarrow m)} = R(\phi)(\boldsymbol{l}_t^{n_i} - \boldsymbol{l}_t^m).$$

In our model, we use a past history of features. The history length (horizon) is denoted as H. The histories of relative gaze directions and relative rotated positions of A^{n_i} w.r.t. A^m are respectively given

by:

$$G^{(n_i \leftarrow m)}_{t,hist} = \left[g^{(n_i \leftarrow m)}_{t-H+1} \quad . \quad . \quad . \quad g^{(n_i \leftarrow m)}_{t} \right] \in \mathbb{R}^{2 \times H},$$
$$L^{(n_i \leftarrow m)}_{t,hist} = \left[l^{(n_i \leftarrow m)}_{t-H+1} \quad . \quad . \quad . \quad l^{(n_i \leftarrow m)}_{t} \right] \in \mathbb{R}^{2 \times H}.$$

The conversational action space of an agent is denoted by $\mathcal{U}$. The conversational actions of A^m and A^{n_i} at time t are represented by $|\mathcal{U}|$-dimensional one-hot vectors d^m_t and $d^{n_i}_t$ respectively. The histories of conversational actions of A^m and A^{n_i} are given by:

$$D^m_{t,hist} = [d^m_{t-H+1}, ..., d^m_t] \in \{1, 0\}^{|\mathcal{U}| \times H}$$
$$D^{n_i}_{t,hist} = [d^{n_i}_{t-H+1}, ..., d^{n_i}_t] \in \{1, 0\}^{|\mathcal{U}| \times H}$$

The state of agent A^m at time t is given by:

$$S^m_t = \left[\left\{ G^{(n_i \leftarrow m)}_{t,hist} \right\}^{J_t}_{i=1}, \left\{ L^{(n_i \leftarrow m)}_{t,hist} \right\}^{J_t}_{i=1}, \left\{ D^{n_i}_{t,hist} \right\}^{J_t}_{i=1}, D^m_{t,hist} \right]$$

Modules. We now present details of our MGpi network's components. As shown in Figure 2, the MGpi network consists of three modules: (a) *Social Signal Gating Module:* This module aggregates observations from surrounding agents in the environment into an internal encoding; (b) *Self Short-Term Memory Encoder:* This encodes each agent's own communication action history; (c) *Interaction Policy:* This module decides the agent's next action based on the agent's encoded observations of its neighbors and of itself. Specifically, the policy incorporates the outputs of the Social Signal Gating Module and Self-STM Encoder and generates a probability vector over the next action of the agent.

3.1 Social Signal Gating Module

Conceptually, the Social Signal Gating (SSG) Module allows the agent to select only the most relevant social signals coming from other agents. Computationally, the SSG generates an encoding for the part of available information that depends on the neighboring agents (the part that depends on itself is described later). As shown in blue in Figure 3, SSG consists of three successive computational units: (1) Social Short-Term Memory Encoder, (2) Kinesic-Proxemic-Message Gate and (3) Signal Pooling.

Social Short-Term Memory Encoder (Social-STM). For agent A^m, at any time t, the role of Social-STM is to generate an encoding for each neighbor A^{n_i}'s social signals. As shown in red in Figure 3, the Social-STM contains two encoders. (1) Non-Verbal Encoder N encodes a history of relative gaze directions and rotated positions $P^{n_i}_{t,hist} = \begin{bmatrix} G^{n_i}_{t,hist} \\ L^{n_i}_{t,hist} \end{bmatrix} \in \mathbb{R}^{4 \times H}$ (2) Conversational Encoder C encodes a history of the neighbor's conversational actions $D^{n_i}_{t,hist}$.

Non-Verbal Encoder $N\left(P^{n_i}_{t,hist}\right)$ summarizes A^m's observations of neighbor A^{n_i}'s non-verbal communication history. Conversational Encoder $C\left(D^{n_i}_{t,hist}\right)$ summarizes A^m's observations of A^{n_i}'s past conversation. As shown in Figure 3, N and C are both linearly-activated gated recurrent units (GRU) [11]. Each encoder learns the associated temporal dynamics and outputs neighbor 'messages' $N \in \mathbb{R}^{64}$ and $C \in \mathbb{R}^{64}$. J_t pairs of messages are output, one pair for each neighbor. Next, these messages are weighted by the KPM gate's importance score, pooled and passed to the interaction policy.

Kinesic-Proxemic-Message (KPM) gate. The role of the KPM gate is to decide how much attention to give to each neighboring agent's concatenated social signal $\left[N\left(P^{n_i}_{t,hist}\right), C\left(D^{n_i}_{t,hist}\right)\right]$. The KPM gate $K\left(g^{(n_i \leftarrow m)}_t, l^{(n_i \leftarrow m)}_t\right)$ is a function of the relative rotated position and orientation between A^m and neighbor A^{n_i} at time t. The KPM gate is a multi-layer perceptron, consisting of two feed-forward layers, respectively activated by an exponential linear unit (ELU) [12] and a hard-sigmoid function [14]. For each neighbor A^{n_i}, it outputs a scalar 'importance' weight $K \in [0, 1]$, reflecting the neighbor's degree of influence on A^m's next conversational action.

Signal Pooling. The role of the signal pooling operator is to aggregate the weighted social signals for all J_t neighbors of A^m at time t. As shown in Figure 3, each neighbor A^{n_i}'s concatenated social signal is weighted by the KPM gate's respective importance score to obtain:

$$x^{(n_i \leftarrow m)}_t = K\left(g^{(n_i \leftarrow m)}_t, l^{(n_i \leftarrow m)}_t\right) \cdot \left[N\left(P^{n_i}_{t,hist}\right) \quad C\left(D^{n_i}_{t,hist}\right)\right]$$

Next, similar to [55], we employ average pooling of neighbors' weighted messages. The pooled message $x^m_t = (1/J_t) \sum_{i=1}^{J_t} x^{(n_i \leftarrow m)}_t$ is passed to the interaction policy π, along with the Self Short-Term Memory Encoder message (described next), to decide A^m's next conversational action.

3.2 Self Short-Term Memory Encoder (Self-STM)

The role of the Self-STM is to encode the agents recollection of her own past actions at time t. Similar to the Conversational Encoder, the Self-STM C' encodes a history of the agent A^m's own conversational actions $D^m_{t,hist}$. As shown in yellow in Figure 3, $C'(D^m_{t,hist})$ is a linearly-activated GRU, yielding a self 'message' $C' \in \mathbb{R}^{64}$. This is concatenated with the pooled neighbor message x^m_t and passed to the interaction policy to decide A^m's next conversational action.

3.3 Interaction Policy

The final module of MGpi is the interaction policy π which, at time t, decides the agent's next action based on the agent's observations of its neighbors and itself. Formally, the policy π takes as input the concatenated outputs $\left[x^m_t, \ C'\left(D^m_{t,hist}\right)\right]$ from the SSG Module and the Self-STM (shown in green in Figures 2, 3). The interaction policy $\pi\left(\left[x^m_t, \ C'\left(D^m_{t,hist}\right)\right]\right)$ is modeled as two fully-connected layers, respectively activated by an ELU and soft-max function. It yields a $|\mathcal{U}|$-dimensional output, representing a probability distribution over the agent A^m's next conversational action.

4 IMITATION LEARNING AND GROUP IDENTIFICATION

By modeling the MGpi Network's modules N, C, G, C', and π as fully-differentiable functions, we ensure that it is trainable end-to-end via back-propagation, using demonstrations of multiagent multigroup state and action sequences [1].

[1] Our code may be found here: https://github.com/navysanghvi/MGpi

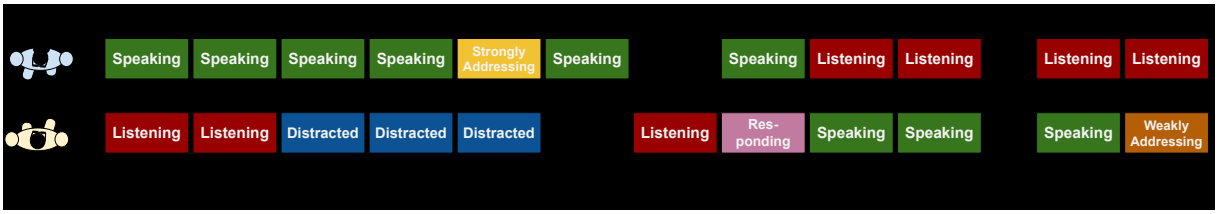

Figure 4: Examples of simulated communications (static two agent case)

Imitation Learning. We perform imitation learning to learn a policy from several 'expert' demonstrations. In this work, we adopt the approach of behavior cloning, which enables MGPI to learn from a set of demonstrations $\mathcal{D}$ directly, via supervised learning without needing to explicitly learn state transitions or reward functions. Similar approaches can be found in recent work on multiagent behavior prediction [1, 45]. We explain how these demonstrations are obtained in Section 5. Each demonstration $D \in \mathcal{D}$ is of length T, and denoted by $D = \{(\boldsymbol{S}_t, \boldsymbol{U}_t)\}_{t=1}^{T}$, where $\boldsymbol{S}_t$ and $\boldsymbol{U}_t$ are the joint state and joint action of the M interacting agents at time t. Therefore, $\boldsymbol{S}_t = \{S_t^m\}_{m=1}^{M}$, and $\boldsymbol{U}_t = \{\boldsymbol{d}_t^m\}_{m=1}^{M}$, where S_t^m and $\boldsymbol{d}_t^m$ are individual agent state and action, as described in the notation in Section 3. Note that $\boldsymbol{S}_t$ contains no information about the group each agent belongs to.

Group Assignments. We emphasize again that training of the MGPI Network requires no supervision with respect to group assignments. Moreover, we identify these assignments implicitly using the KPM gate module K. For every agent A^m, in order to minimize the error between MGPI's predicted action $\hat{\boldsymbol{d}}_t^m$ and demonstrated action $\boldsymbol{d}_t^m$, KPM gate K must implicitly learn spatial attention. The higher the 'importance' weight assigned by K to a neighbor's encoded non-verbal and conversational history, the more likely the neighbor is to be in the same group as the A^m, and the greater its influence on A^m's resulting policy π.

Group Identification. Once the MGPI network parameters are learned, we use the KPM gate activations to identify groups. In order to identify groups, at each time step, we define a distance $D(n, j)$ between two agents n, j based on the output of learned gating function K: $D(n, j) = 1 - \frac{1}{2}(K(\boldsymbol{g}^{(j \leftarrow n)}, \boldsymbol{l}^{(j \leftarrow n)}) + K(\boldsymbol{g}^{(n \leftarrow j)}, \boldsymbol{l}^{(n \leftarrow j)}))$. The distance is simply computed from the weighted average of the bi-directional weights computed by the KPM gate. We compose an affinity matrix of pairwise distances between all agents and run the DBSCAN clustering algorithm [22] to cluster people into conversational groups. As mentioned in our Introduction, this is a relatively simple and straight-forward approach. It is compared in Section 6 with state-of-the-art approaches based on complex *o-space* estimation under assumptions of particular spatial layouts called F-formations [34].

5 EXPERIMENTS: SOCIAL INTERACTION SIMULATOR

Conversational Action Data. We aim to learn computational models that can encode the kinesics and proxemics in group formation and conversational actions (*e.g.*, responding, distraction, *etc*) in multiagent multigroup scenarios. Unfortunately, there are no public datasets of large-scale demonstrations of the physical

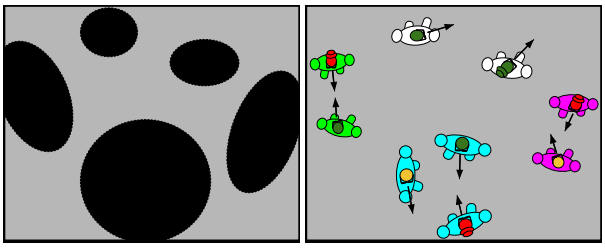

Figure 5: Example physical layout [15] and subsequent conversational interaction. Arrows represent gaze directions, head colors indicate conversational action - red: speaking, yellow: distracted, green: listening.

positions, gaze directions, group assignments, and communication actions of multiple socially interacting people. Existing datasets [15, 45, 62] only provide sequences of the physical positions, gaze directions and group assignments (but annotated only sparsely). More importantly, *no conversational action annotations are provided.* Similarly, there are no social interaction simulators for the kind of face-to-face behavioral actions of interest to us. In order to evaluate the capacity of our proposed model, we simulate multigroup human communication, drawing inspiration from several prior studies and observations of small-group conversational dynamics [27, 28, 30, 33, 40]. While we simulate sequences of non-verbal interactions, we point out that the physical layouts of agents are from real-world datasets.

Initial Agent Layout Data: We experiment with both artificially constructed and real-world data to form our initial multigroup layouts - specifically, we use (a) Synthetic [15][2], (b) Coffee Break [15], and (c) Cocktail Party [62] datasets to provide 100, 119, and 320 non-identical, independent layouts respectively. Multiple people (6-12 per layout) are organized into several groups (Fig. 1).

Rules of Interaction: To design multigroup communication protocols, we draw inspiration from several observations of small-group conversational dynamics in literature, namely: speaking and turn-taking, [27], addressing [33], interest in meetings [28, 30], and the relation of speaking time and dominance [40]. We design per-group rules for two scenarios: **(1) Static scenario**, where agents stay within a social group and do not transition to other groups. Six conversational actions are considered $\mathcal{U}^{stat} = \{$*Speaking*, *Listening*, *Distracted*, *Strongly Addressing*, *Weakly Addressing*, *Responding*$\}$. **(2) Dynamic scenario**, where agents transition from group to group, leading to a temporally evolving group assignments. Seven conversational actions are considered $\mathcal{U}^{dyn} = \mathcal{U}^{stat} \cup \{$*Moving*$\}$.

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

Table 1: Results for Static Scenario: Action Prediction (mean average precision), and Test Loss (cross-entropy)

Model **(mAP's)**	$J = 2$	4	8	12
Neighbor States Only (NSO)	0.67	0.69	0.70	0.68
Self State Only (SSO)	0.76	0.76	0.76	0.76
Equal Pooling (EQPOOL)	0.87	0.86	0.85	0.85
Social Pooling (SOCPOOL)	0.85	0.84	0.84	0.84
MGpi Network	**0.88**	**0.88**	**0.89**	**0.88**

Model **(cross-entropy losses)**	$J = 2$	4	8	12
Neighbor States Only (NSO)	0.99	0.95	0.89	1.00
Self State Only (SSO)	0.29	0.29	0.29	0.29
Equal Pooling (EQPOOL)	0.21	0.22	0.22	0.22
Social Pooling (SOCPOOL)	0.21	0.22	0.22	0.25
MGpi Network	**0.20**	**0.21**	**0.20**	**0.21**

Table 2: Results for Dynamic Scenario: Action Prediction (mean average precision), and Test Loss (cross-entropy)

Model **(mAP's)**	$J = 2$	4	8	12
Neighbor States Only (NSO)	0.57	0.62	0.65	0.64
Self State Only (SSO)	0.78	0.78	0.78	0.78
Equal Pooling (EQPOOL)	0.85	0.86	0.85	0.85
Social Pooling (SOCPOOL)	0.85	0.85	0.86	0.85
MGpi Network	**0.87**	**0.87**	**0.88**	**0.88**

Model **(cross-entropy losses)**	$J = 2$	4	8	12
Neighbor States Only (NSO)	1.24	1.12	1.05	1.07
Self State Only (SSO)	0.31	0.31	0.31	0.31
Equal Pooling (EQPOOL)	0.26	0.26	0.25	0.25
Social Pooling (SOCPOOL)	0.27	0.27	0.25	0.26
MGpi Network	**0.25**	**0.23**	**0.25**	**0.22**

Based on observations from literature described above, we design per-group probabilistic rules conditioned on a temporal history of past group actions, to decide the evolution of each agent's conversational actions. The same rules are applied to every group in the scene. We simulate 600-step interaction episodes starting from each set of initial layouts ((a),(b) and (c) above) to obtain training data. Figure 5 shows an example of initial multigroup layout and subsequent interaction[3].

We reiterate: group assignment information is used *only for evaluation* and *not* for supervisory training signals to MGpi. Instead, MGpi implicitly identifies groups using the KPM gate. We hypothesize that our automatic KPM gating mechanism will learn to discover relevant neighbors in both static and dynamic scenarios. While we recognize the gap between reality and our simulations, we believe that, in the absence of existing data or other simulators, this is a necessary first step towards understanding how to model multiagent multigroup interaction.

Two-Agent Simulation Example. Using Figure 4, we illustrate our rules in a simple two-agent case. In the static scenario, as shown in Figure 4(a), one agent is *Speaking* while the other is *Listening*. The *Listening* agent may become *Distracted* (Figure 4(b)) with a certain probability. After some time steps, the *Speaking* agent may start *Strongly Addressing* the group, to draw back the attention of the *Distracted* member (Figure 4(c)). Once the *Speaking* agent has spoken for some time, it yields to the other. The other agent is then *Responding* followed by *Speaking* (Figure 4(d)). The *Speaking* agent might transition to *Weakly Addressing* (Figure 4(e)). Additionally, in a dynamic scenario, a group member who is *Distracted* may start *Moving* toward another group. When a *Moving* agent joins a new group, it is welcomed as the next speaker.

Baselines. We now describe experiments and results of applying the MGpi network to predict communication actions of simulated social agents. We compare MGpi with the following baselines, including communication models used in prior work:

[3]Example simulator videos may be found here: https://github.com/navysanghvi/MGpi

Neighbor State Only (NSO): We omit Self-STM Encoder C' so as to only use cues from neighbors. This helps us test whether observations of the self are necessary for modeling social interaction.

Self State Only (SSO): We omit the Social Signal Gating Module so as to use cues of the agent's self alone. This helps us test whether observations of nearby agents' social signals are necessary for modeling social interaction.

Equal Pooling (EQPOOL): Prior work studies the role of message broadcasting in scenarios where one agent simply averages messages it receives from other agents [24, 43, 55]. We implement this message broadcasting baseline by omitting the KPM gate K so that all neighbor encoded 'messages' N and C are *equally weighted*, such that $\boldsymbol{x}_t^{(n_i \leftarrow m)} = \left[N\left(\boldsymbol{P}_{t,hist}^{n_i}\right) \;\; C\left(\boldsymbol{D}_{t,hist}^{n_i}\right)\right]$. This helps us test if equal social perception to neighbor agents is sufficient for modeling social interaction.

Social Pooling (SOCPOOL): We adopt social pooling [1, 37] as a second message broadcasting baseline. It uses a grid that divides the 2D world into non-overlapping regions and pools messages by region. In our experiments, we use a 4×4 grid around target locations, each of size 50×50 pixels (this configuration yielded best performance). For each region, we average $\boldsymbol{x}_t^{(n_i \leftarrow m)}$ over neighbors A^{n_i} located in that region. For this baseline, the KPM gate and original Signal Pooling Mechanism of MGpi are replaced by an operator that performs grid-wise pooling and concatenates pooled messages. We expect this baseline to learn to implicitly gate messages by down-weighting messages in far away grids.

Training/Evaluation Scheme. Demonstrations of multigroup communication were collected as described previously. For each layout in each dataset (Synthetic, Coffee Break and Cocktail Party), agents changed modes 600 times, *i.e.*, $T = 600$. We split demonstrations into two subsets with equal number of layouts from each dataset, and learned models on one subset to test the other. All metrics are averaged over the two test subsets for two-fold cross-validation. All models were trained for 30 epochs with mini-batches of size 4096, various numbers of agents $J \in \{2, 4, 8, 12\}$, and a history window $H = 15$, and we confirmed that neither further

training nor smaller mini-batches improved performance. Each network was trained to minimize categorical cross entropy (also called negative log likelihood) between predicted probabilities of actions $\{\{\hat{U}_t\}_{t=1}^{T}\}_{\mathcal{D}}$ and actual demonstrated actions $\{\{U_t\}_{t=1}^{T}\}_{\mathcal{D}}$ via Adam [35]. We report the cross-entropy loss on testing subsets. As every model assigns maximum probability to one of six (static scenario) or seven (dynamic scenario) actions, we also evaluate the model on mean average precision (mAP) over actions.

Results. Tables 1 and 2 show all model performances in communication action prediction, with $H = 15$, and various numbers of neighbors J. We emphasize that J is only fixed during training; since a single Social-STM Encoder is learned, *any* number of neighbors may be considered at every time step during testing - in fact, *different* numbers of neighbors may be considered for each agent in the scene.

As hypothesized, MGPI effectively learns protocols in both static and dynamic scenarios, outperforming strong baselines in terms of both mean average precision (mAP) and test loss, validating our choices of various encoding modules and demonstrating each one's necessity. Confusion matrices of action prediction for all networks (omitted here for spcae) show that NSO and SSO work complementarily: NSO predicts better if agents make strong or weak addressing decisions by observing others. On the other hand, SSO performs better when observation of the self is necessary, *i.e.*, in order to predict if agents keep speaking, listening, are distracted or responding. MGPI outperforms baselines for most actions, both in the static and dynamic case, with lower ambiguity.

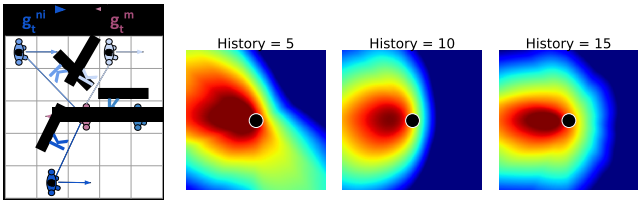

Figure 6: Social Attention of KPM Gate: For various positions of a neighbor A^{n_i} (*e.g.*, any blue figure) in a fine grid around A^m (magenta figure, black circles), warmer colors indicate higher outputs from the learned KPM gate K. Gaze directions $g_t^m = [-1, 0]$, $g_t^{n_i} = [+1, 0]$. History windows $H \in \{5, 10, 15\}$.

Learned Social Attention. Figure 6 represents physical constraints learned by the MGPI Network's KPM gate K to judge if neighbor agent A^{n_i} is in the same group as agent A^m. These visuals are obtained after training in the static scenario. We visualize the output from K at various locations around A^m while fixing relative gaze $g^{(n_i \leftarrow m)}$. As expected, we see improved learning with larger history windows H, and choose $H = 15$ in our experiments. We confirm that increasing the history window does not improve performance, suggesting that, once trained on real-world scenarios, MGPI might be able to *reflect the optimal length of short-term memory exhibited in humans.* Also as expected, we see a higher output from our KPM gate when agents are looking toward each other, i.e., when the neighbor A^{n_i} is toward the left of the image, looking in $+x$ direction, *i.e.*, toward A^m, who is looking in $-x$ direction.

6 EXPERIMENTS: GROUP IDENTIFICATION IN REAL DATA

We have demonstrated the effectiveness of the MGPI network in learning group communication policy with simulated social agents. We now test the effectiveness of our learned automatic social signal gating mechanism, the KPM gate, in identifying communication groups in real-world data [15, 62].

Training MGPI. Similar to Section 5, we train MGPI using a dataset of simulated interactions from a mixture of initial layouts from the Synthetic [15], Coffee Break [15] and Cocktail Party [62] datasets. For the sake of equal representation in our training data, we use 100 initial layouts from each dataset, for a total of 300 demonstrations of $T = 600$ steps each. Each dataset has several continuous frames with various numbers of people annotated with positions, orientations, and group assignments (Fig. 1). MGPI is trained for 20 epochs, in both static and dynamic scenarios, setting $J = 12$, which is sufficiently larger than group membership (6 at most) in the datasets. This ensures that the network must learn to filter out irrelevant social cues from non-group agents, strengthening the contribution of the KPM gate to good network performance. Further, this enhances the chances that KPM gate's learned social attention is effective for group identification.

Unsupervised Group Identification. As described in Section 4, once MGPI is trained, we define a distance $D(n, j)$ between two agents n, j based on the output of learned gating function K: $D(n, j) = 1 - \frac{1}{2}(K(\boldsymbol{g}^{(j \leftarrow n)}, \boldsymbol{l}^{(j \leftarrow n)}) + K(\boldsymbol{g}^{(n \leftarrow j)}, \boldsymbol{l}^{(n \leftarrow j)}))$. For every initial layout in real-world datasets Synthetic, Coffee Break, and Cocktail Party [15, 62], we use this simply computed distance measure to form an affinity matrix of pairwise distances between agents in the scene. Using this affinity matrix, we run the DBSCAN clustering algorithm [22] to cluster people into conversational groups.

Evaluation. Group detection performance is measured by the F1 score (harmonic mean of Precision and Recall) under the $|G|$ condition [15, 50]. Per frame, a group is judged to be detected correctly if *all* constituent people are grouped into a single cluster. Precision, Recall and F1 scores are averaged over frames for each dataset (Synthetic, Coffee Break, Cocktail Party) separately, as well as over all datasets. Due to the stochastic nature of training MGPI, we show average and standard deviation of KPM gate's performance over 5 training runs.

Comparisons. We define a pose-only baseline, which runs DBSCAN using actual euclidean distances between agents in each scene, scaled by the maximum such distance. Unlike MGPI and other state-of-the-art methods, this does not use gaze or body orientation information. In addition to this simple baseline, several state-of-the-art group identification methods DS[32], HVFF-ms [50], GCFF [51], GRUPO [58] are chosen for comparison. These methods are based on complex *o-space* estimation under assumptions of particular spatial layouts called F-formations [34], often using heuristic parameters. By contrast, KPM gate offers a direct, unsupervised approach to group identification.

Results. From Table 3 we see that our learning process of the KPM gate is very stable and has very low standard deviation across training runs. Our simple method takes a few miliseconds to run and could be used for real-time group identification. On the Synthetic dataset, our approach significantly outperforms all state-of-the-art

Table 3: Evaluation on group detection: Precision, Recall and F1 scores under the $|G|$ criterion. State-of-the-art methods DS [32], HVFF-ms [50], GCFF [51], and GRUPO [58] are compared against a simple baseline and the KPM gate (ours). Standard deviation across 5 training runs are shown in brackets. Due to averaging over a lesser number of datasets, italicized results for DS and GRUPO cannot be taken into account for a fair comparison

Method⇒		DS	HVFF-ms	GCFF	GRUPO	Pose-only	KPM gate (ours)	
Dataset ⇓	Metric ⇓					Baseline	Static Scenario	Dynamic Scenario
Synthetic	Precision	0.68	0.72	0.91	N/A	0.88	**1.00** (0.000)	**1.00** (0.000)
	Recall	0.80	0.73	0.91	N/A	0.58	**0.96** (0.008)	**0.98** (0.000)
	F1 Score	0.74	0.73	0.91	N/A	0.70	**0.98** (0.004)	**0.99** (0.000)
Coffee Break	Precision	0.40	0.40	0.61	N/A	0.53	**0.63** (0.004)	**0.61** (0.004)
	Recall	0.38	0.38	**0.64**	N/A	0.46	0.63 (0.010)	0.62 (0.005)
	F1 Score	0.39	0.39	**0.63**	N/A	0.49	**0.63** (0.005)	0.62 (0.001)
Cocktail Party	Precision	N/A	0.30	0.63	**0.65**	0.29	0.60 (0.004)	0.63 (0.010)
	Recall	N/A	0.30	**0.65**	0.63	0.27	0.56 (0.007)	0.55 (0.003)
	F1 Score	N/A	0.30	**0.64**	**0.64**	0.28	0.58 (0.005)	0.59 (0.005)
Overall Mean	Precision	*0.54*	0.47	0.71	*0.65*	0.57	**0.74** (0.002)	**0.75** (0.004)
	Recall	*0.59*	0.47	**0.73**	*0.63*	0.44	0.72 (0.008)	0.72 (0.002)
	F1 Score	*0.56*	0.47	0.72	*0.64*	0.49	**0.73** (0.004)	**0.73** (0.002)

methods in terms of all metrics. On the Coffee Break dataset, our approach outperforms other methods in terms of Precision. and performs as well as the best (GCFF) in terms of F1 score. On the Cocktail Party dataset, our method performs slightly worse than the best (GCFF and GRUPO), but better than HVFF-ms. Overall, our method outperforms all methods in terms of Precision and F1 score, and performs comparably to the best (GCFF) in terms of Recall.

Discussion. Our work improves over prior work in ways that are even more significant than our superior overall performance:
(1) Imitation leads to group detection: Most importantly, unlike prior work, our model is not trained explicitly for group detection. Instead, our focus is on behavioral imitation (learning a model that mimics human interaction). The surprising result is that, after training for behavior imitation, the output of KPM gate also performs successfully on the task of group detection.
(2) Data-driven approach, features, parameters: All methods we compare with rely on the assumption that people in groups position themselves in specific layouts (F-formations) developed in interactional socio-linguistics. By contrast, our models use no prior knowledge about layouts - we learn spatial layouts from data. Methods like GRUPO [58] use hand-defined parameters, features (e.g, stride, lambda, gaussian mixtures), and expensive iterative mode-finding sub-algorithms. By contrast, our method has a single learned feature (KPM gate) and performs low-cost O(N^2) unsupervised clustering. GRUPO uses the additional feature of lower body orientation, which our method does not.

In summary, we have demonstrated state-of-the-art performance using a simple, direct method without the use of hand-crafted parameters, features or layout assumptions. The remarkable *emergence* of social perception in the form of group identification has been compared with various *explicit* methods for this task.

7 CONCLUSION

In this work, we presented MGpi, a deep neural network model of non-verbal and conversational interactions among multiple people and multiple groups. In the absence of real-world annotated data or simulators of multigroup face-to-face conversational behavior (*e.g.*, speaking, listening, responding), we have designed our own social interaction simulator. While we consider discrete conversational actions among agents, we believe this is a necessary first step toward designing more complex models of real-world multigroup social intelligence.

We have demonstrated the necessity and efficacy of each interpretable component of MGpi, by evaluating several ablative baselines on their ability to predict the next appropriate action. We have demonstrated MGpi's superior performance in scenarios with static and dynamically evolving group assignments. We reiterate that training MGpi requires no explicit group annotations. Instead, its Kinesic-Proxemic-Message gate (KPM gate) learns to express social attention based on non-verbal cues, as a result of training for a socially intelligent policy.

We have demonstrated the remarkable emergence of KPM gate's ability to identify groups in real-world data, and achieved state-of-the-art results with our direct, unsupervised method, without using complex layout assumptions or hand-defined parameters.

In a dearth of real-world data on multiagent multigroup conversational interactions, valuable future work would involve large-scale collection of non-verbal and verbal exchanges between multiple groups of people. The availability of more diverse data of human-to-human exchange (*e.g.*, body orientation, gestures, facial expressions, *etc*) would equip us to better design computational models like MGpi. Furthermore, such data would equip our models to better reason about social perception and propriety, by allowing the training of deep modules for a wider range of classes of behavioral actions, *e.g.*, illustrators, emblems and attitudes [59].

Acknowledgment: This project was sponsored in part by JST CREST (JPMJCR14E1), NSF NRI (1637927) and IARPA (D17PC00340).

REFERENCES

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. 2016. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*.

[2] Karl Albrecht. 2006. *Social intelligence: The new science of success*. John Wiley & Sons.

[3] Loris Bazzani, Marco Cristani, Diego Tosato, Michela Farenzena, Giulia Paggetti, Gloria Menegaz, and Vittorio Murino. 2013. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems* 30, 2 (2013), 115–127.

[4] Ray L Birdwhistell. 1952. *Introduction to kinesics: An annotation system for analysis of body motion and gesture*. University of Louisville.

[5] Manuele Brambilla, Eliseo Ferrante, Mauro Birattari, and Marco Dorigo. 2013. Swarm robotics: a review from the swarm engineering perspective. *Swarm Intelligence* 7, 1 (2013), 1–41.

[6] Cynthia Breazeal. 2003. Toward sociable robots. *Robotics and autonomous systems* 42, 3-4 (2003), 167–175.

[7] Cynthia Breazeal, Daphna Buchsbaum, Jesse Gray, David Gatenby, and Bruce Blumberg. 2005. Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. *Artificial life* 11, 1-2 (2005), 31–62.

[8] Adelbert W Bronkhorst. 2000. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica* 86, 1 (2000), 117–128.

[9] L. Busoniu, R. Babuŝka, and B. D. Schutter. 2008. A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 2 (March 2008), 156–172. https://doi.org/10.1109/TSMCC.2007.913919

[10] Ginevra Castellano, Iolanda Leite, Andre Pereira, Carlos Martinho, Ana Paiva, and Peter W McOwan. 2012. Detecting engagement in HRI: An exploration of social and task-based context. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 421–428.

[11] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *CoRR* abs/1412.3555 (2014).

[12] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *CoRR* abs/1511.07289 (2015).

[13] Francisco L Colino, Gavin Buckingham, Darian T Cheng, Paul van Donkelaar, and Gordon Binsted. 2014. Tactile gating in a reaching and grasping task. *Physiological reports* 2, 3 (2014).

[14] Matthieu Courbariaux and Yoshua Bengio. 2016. BinaryNet: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1. *CoRR* abs/1602.02830 (2016).

[15] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. 2011. Social interaction discovery by statistical analysis of F-formations. In *British Machine Vision Conference*. 23.1–23.12.

[16] Marco Cristani, Vittorio Murino, and Alessandro Vinciarelli. 2010. Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 51–58.

[17] Marco Cristani, Ramya Raghavendra, Alessio Del Bue, and Vittorio Murino. 2013. Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing* 100 (2013), 86–97.

[18] Howard C Cromwell, Ryan P Mears, Li Wan, and Nash N Boutros. 2008. Sensory gating: a translational effort from basic to clinical science. *Clinical EEG and Neuroscience* 39, 2 (2008), 69–72.

[19] Kerstin Dautenhahn. 2007. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences* 362, 1480 (2007), 679–704.

[20] Martin Davies and Tony Stone. 1995. Mental Simulation: Evaluations and Applications-Reading in Mind and Language. (1995).

[21] T Edward. 1966. Hall, The Hidden Dimension.

[22] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*. 226–231.

[23] Stephen M Fiore, Travis J Wiltshire, Emilio JC Lobato, Florian G Jentsch, Wesley H Huang, and Benjamin Axelrod. 2013. Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior. *Frontiers in psychology* 4 (2013), 859.

[24] Jakob Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. In *Proceeding of the Advances in Neural Information Processing Systems*. 2137–2145.

[25] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2017. Counterfactual Multi-Agent Policy Gradients. *CoRR* abs/1705.08926 (2017). http://arxiv.org/abs/1705.08926

[26] Robert Freedman, Lawrence E Adler, Greg A Gerhardt, Merilyne Waldo, Neil Baker, Greg M Rose, Carla Drebing, Herbert Nagamoto, Paula Bickford-Wimer, and Ronald Franks. 1987. Neurobiological studies of sensory gating in schizophrenia. *Schizophrenia bulletin* 13, 4 (1987), 669–678.

[27] D. Gatica-Perez. 2006. Analyzing Group Interactions in Conversations: a Review. In *Proceedings of the International Conference on Multisensor Fusion and Integration for Intelligent Systems*. 41–46.

[28] Daniel Gatica-Perez, L McCowan, Dong Zhang, and Samy Bengio. 2005. Detecting group interest-level in meetings. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Vol. 1. IEEE, I–489.

[29] Daniel Goleman. 2007. *Social intelligence*. Random house.

[30] Dirk Heylen, Dennis Reidsma, and Roeland Ordelman. 2006. Annotating state of mind in meeting data. In *The Workshop Programme Corpora for Research on Emotion and Affect*. 84.

[31] SA Hillyard and GR Mangun. 1987. Sensory gating as a physiological mechanism for visual selective attention. *Electroencephalography and clinical neurophysiology. Supplement* 40 (1987), 61–67.

[32] Hayley Hung and Ben Kröse. 2011. Detecting F-formations As Dominant Sets. In *Proceedings of the International Conference on Multimodal Interfaces*. 231–238.

[33] Michael Katzenmaier, Rainer Stiefelhagen, and Tanja Schultz. 2004. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 144–151.

[34] A. Kendon. 1990. *Conducting interaction: Pattern of behavior in focused encounter*. Cambridge University Press.

[35] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). http://arxiv.org/abs/1412.6980

[36] Hoang M. Le, Yisong Yue, and Peter Carr. 2017. Coordinated Multi-Agent Imitation Learning. In *Proceeding of the International Conference on Machine Learning*.

[37] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B. Choy, Philip H. S. Torr, and Manmohan Chandraker. 2017. DESIRE: Distant Future Prediction in Dynamic Scenes With Interacting Agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[38] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *CoRR* abs/1706.02275 (2017). http://arxiv.org/abs/1706.02275

[39] Stacy C Marsella, David V Pynadath, and Stephen J Read. 2004. PsychSim: Agent-based modeling of social interactions and influence. In *Proceedings of the international conference on cognitive modeling*, Vol. 36. 243–248.

[40] Marianne Schmid Mast. 2002. Dominance as expressed and inferred through speaking time: A meta-analysis. *Human Communication Research* 28, 3 (2002), 420–450.

[41] Albert Mehrabian and Susan R Ferris. 1967. Inference of attitudes from nonverbal communication in two channels. *Journal of consulting psychology* 31, 3 (1967), 248.

[42] Andrew N Meltzoff and Jean Decety. 2003. What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 358, 1431 (2003), 491–500.

[43] Igor Mordatch and Pieter Abbeel. 2017. Emergence of Grounded Compositional Language in Multi-Agent Populations. *CoRR* abs/1703.04908 (2017). http://arxiv.org/abs/1703.04908

[44] Peng Peng, Quan Yuan, Ying Wen, Yaodong Yang, Zhenkun Tang, Haitao Long, and Jun Wang. 2017. Multiagent Bidirectionally-Coordinated Nets for Learning to Play StarCraft Combat Games. *CoRR* abs/1703.10069 (2017). http://arxiv.org/abs/1703.10069

[45] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. 2016. Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes. In *Proceeding of the European Conference on Computer Vision*. 549–565.

[46] Hanan Salam, Oya Celiktutan, Isabelle Hupont, Hatice Gunes, and Mohamed Chetouani. 2016. Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. *IEEE Access* 5 (2016), 705–721.

[47] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th international conference on Human-robot interaction*. ACM, 305–312.

[48] Navyata Sanghvi, Sasanka Nagavalli, and Katia Sycara. 2017. Exploiting robotic swarm characteristics for adversarial subversion in coverage tasks. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 511–519.

[49] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[50] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani. 2013. Multi-scale f-formation discovery for group detection. In *Proceedings of the International Conference on Image Processing*. 3547–3551.

[51] Francesco Setti, Chris Russell, Chiara Bassetti, and Marco Cristani. 2015. F-Formation Detection: Individuating Free-Standing Conversational Groups in Images. *PLOS ONE* 10, 5 (2015), 1–26.

[52] Kimron L Shapiro, Judy Caldwell, and Robyn E Sorensen. 1997. Personal names and the attentional blink: A visual" cocktail party" effect. *Journal of Experimental Psychology: Human Perception and Performance* 23, 2 (1997), 504.

[53] Satinder P Singh, Michael J Kearns, Diane J Litman, and Marilyn A Walker. 2000. Reinforcement learning for spoken dialogue systems. In *Advances in Neural Information Processing Systems*. 956–962.

[54] Peter Stone and Manuela Veloso. 2000. Multiagent Systems: A Survey from a Machine Learning Perspective. *Autonomous Robots* 8, 3 (01 Jun 2000), 345–383. https://doi.org/10.1023/A:1008942012299

[55] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2016. Learning Multiagent Communication with Backpropagation. In *Proceeding of the Advances in Neural Information Processing Systems*. 2244–2252.

[56] Mason Swofford, John Charles Peruzzi, Marynel Vázquez, Roberto Martín-Martín, and Silvio Savarese. 2019. DANTE: Deep Affinity Network for Clustering Conversational Interactants. *arXiv preprint arXiv:1907.12910* (2019).

[57] Sebastiano Vascon, Eyasu Zemene Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. 2014. A game-theoretic probabilistic approach for detecting conversational groups. In *Asian conference on computer vision*. Springer, 658–675.

[58] Marynel Vazquez, Aaron Steinfeld, and Scott E. Hudson. 2015. Parallel Detection of Conversational Groups of Free-Standing People and Tracking of their Lower-Body Orientation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*.

[59] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing* 27, 12 (2009), 1743–1759.

[60] Alan R Wagner and Ronald C Arkin. 2011. Acting deceptively: Providing robots with the capacity for deception. *International Journal of Social Robotics* 3, 1 (2011), 5–26.

[61] Woo-Han Yun, Dongjin Lee, Chankyu Park, Jaehong Kim, and Junmo Kim. 2018. Automatic Recognition of Children Engagement from Facial Video using Convolutional Neural Networks. *IEEE Transactions on Affective Computing* (2018).

[62] Gloria Zen, Bruno Lepri, Elisa Ricci, and Oswald Lanz. 2010. Space Speaks: Towards Socially and Personality Aware Visual Surveillance. In *Proceeding of the International Workshop on Multimodal Pervasive Video Analysis*. 37–42.