Group-based Motion Prediction for Navigation in Crowded Environments

Anonymous Author(s)
Affiliation
Address
email

Abstract: We focus on the problem of planning the motion of a robot in a dynamic multiagent environment such as a pedestrian scene. Enabling the robot to navigate safely and in a socially compliant fashion in such scenes requires a representation that accounts for the unfolding multiagent dynamics. Existing approaches to this problem tend to employ microscopic models of motion prediction that reason about the individual behavior of other agents. While such models may achieve high tracking accuracy in trajectory prediction benchmarks, they often lack an understanding of the group structures unfolding in crowded scenes. Inspired by the Gestalt theory from psychology, we build a Model Predictive Control framework (G-MPC) that leverages group-based prediction for robot motion planning. We conduct an extensive simulation study involving a series of challenging navigation tasks in scenes extracted from two real-world pedestrian datasets. We illustrate that G-MPC enables a robot to achieve statistically significantly higher safety and lower number of group intrusions than a series of baselines featuring individual pedestrian motion prediction models. Finally, we show that G-MPC can handle noisy lidar-scan estimates without significant performance losses.

1 Introduction

Over the past three decades, there has been a vivid interest in the area of robot navigation in pedestrian environments [1, 2, 3, 4, 5]. Planning robot motion in such environments can be challenging due to the lack of rules regulating traffic, the close proximity of agents and the complex emerging multiagent interactions. Further, accounting for human safety and comfort as well as robot efficiency add to the complexity of the problem.

To address such specifications, a common [6, 4, 3, 7, 8] paradigm involves the integration of a behavior prediction model into a planning mechanism. Recent models tend to predict the individual interactions among agents to enable the robot to determine collision-free candidate paths [3, 4, 9]. While this paradigm is well-motivated, it tends to ignore the structure of interaction in such environments. Often, the motion of pedestrians is coupled as a result of social grouping. Further, the motion of multiple agents can often be *effectively* grouped as a result of similarity in motion characteristics. Lacking a mechanism for understanding the emergence of this structure, the robot motion generation mechanism may yield unsafe or uncomfortable paths for human bystanders, often violating the space of social groups.

Motivated by such observations, we draw inspiration from human navigation to propose the use of group-based prediction for planning in crowd navigation domains. We argue that humans do not employ detailed individual trajectory prediction mechanisms. In fact, our motion prediction capabilities are short-term and do not scale with the number of agents. We do however employ effective grouping techniques that enable us to discover safe and efficient paths among motions of crowd networks. This anecdotal observation is aligned with gestalt theory from psychology [10] which suggests that organisms tend to perceive and process *formations of entities*, rather than individual components. Such techniques have recently led to advances in computer vision [11] and computational photography [12]. Similarly, we envision that a robot could reason the formation of effective groups in a crowded environment and react to their motion as an effective way to navigate safely.

Submitted to the 5th Conference on Robot Learning (CoRL 2021). Do not distribute.

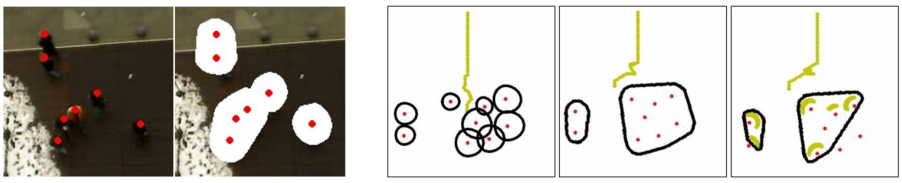

Figure 1: Based on a representation of social grouping [13], we build a group behavior prediction model to empower a robot to perform safe and socially compliant navigation in crowded spaces. The images to the left demonstrate an example of our representation overlayed on top of a scene from a real-world dataset [14]. The images to the right demonstrates that a model predictive controller equipped with our prediction model is able to navigate around the group socially (middle) as opposed to the baseline that cuts through the group (left). Our formulation is also able to handle imperfect state estimates (right) where the yellow arcs are scan points from a simulated 2D lidar laser scan.

In this paper, we propose a group-based representation coupled with an autoencoder prediction model based on the group-space approximation model of Wang and Steinfeld [13]. This model groups a crowd into sets of agents with similar motion characteristics and draws geometric enclosures around them, given observation of their states. The prediction module then predicts future states of these enclosures. We conduct an extensive empirical evaluation over 5 different human datasets [14, 15], each with a flow following and a crossing scenario. We further conduct a same set of evaluations with agents powered by ORCA [16] that share the start and end locations in the datasets. Last but not least, we conducted evaluation given inputs in the form of simulated laser scans, from which pedestrians are only partially observable or even completely occluded. We compare the performance of our group-based formulation against three individual reasoning baselines: a) a reactive baseline with no prediction; b) a constant velocity prediction baseline; c) one based on individual S-GAN trajectory predictions [17]. We present statistically significant evidence suggesting that agents powered by our formulation produce safer and more socially compliant behavior and are potentially able to handle imperfect state estimates.

2 Related Work

Over the recent years, a considerable amount of research has been placed to the problem of robot navigation in crowded pedestrian environments [4, 3, 18, 8, 19, 20, 21, 22, 23]. Such environments often comprise groups of pedestrians, navigating as coherent entities. This has motivated recent work on group detection and group motion modeling.

Groups are often perceived as sub-modular entities that collectively define the behavior of the crowd. Šochman and Hogg [24] suggests that 50-70% of pedestrians walk in groups. Many works exist in group detection. One popular area in such domain is static group detection, often leveraging F-formation theories [25]. However, dynamic groups often dominate pedestrian rich environments and they exhibit different spatial behavior [26]. Among dynamic group detection, the most common approach treats grouping as a probabilistic process where groups are a reflection of close probabilistic association of pedestrian trajectories [27, 28, 29, 30, 31]. Others use graph models to build inter-pedestrian relationships with strong graphical connections indicating groups [32, 33]. The social force model [34] also inspires Mazzon et al. [35], Šochman and Hogg [24] to develop features that indicate groups. Clustering is another common group of technique to group pedestrians with similar features into groups [36, 37, 38, 39]. In this paper, we do not intend to explore the state-of-the-art grouping practice. For our formulation, it is sufficient to employ a simple clustering based grouping method proposed by Chatterjee and Steinfeld [39].

Applications on groups often focus on a specific behavior aspect. In terms of interacting with pedestrians, a major focus in this area is how a robot should behave as part of the group formation [40]. On dyad groups involving a single human and a robot, some researchers examined socially appropriate following behavior [41, 42, 43, 44] and conversely, guiding behavior [45, 46, 47]. In works that do not include robots as part of the pedestrian groups, some research teams studied how a robot should guide a group of pedestrians [48, 49, 50]. From navigation perspective, Yang and Peters [26] leverage groups as obstacles, but their group space is a collection of individual personal spaces with occasional O-space modeling from F-formation theories. Without the engineered occasional occurrence of O-space, their representation reduces to one of our baselines. Katyal et al. [51] introduce an

additional cost term that leverages robot's distance to the closest group in a reinforcement learning framework. They model groups using convex hulls directly generated from pedestrian coordinates instead of taking personal spaces into consideration. This less principled approach often leads to the robot approaching dangerously close to pedestrians. In our work, we additionally explore the capabilities of groups in handling imperfect sensor inputs. While our focus is on analysing the benefits of groups, our group based formulation can be easily incorporated into Katyal et al. [51]'s framework.

3 Problem Statement

Consider a robot navigating in a workspace $\mathcal{W} \subseteq \mathbb{R}^2$ amongst n other dynamic agents. Denote by $s \in \mathcal{W}$ the state of the robot and by $s^i \in \mathcal{W}$ the state of agent $i \in \mathcal{N} = \{1, \ldots, n\}$. The robot is navigating from a state s_0 towards a destination s_T by executing a policy $\pi : \mathcal{W}^{n+1} \times \mathcal{U} \rightarrow \mathcal{U}$ that maps the assumed fully observable world state $\boldsymbol{S} = s \cup_{i=1:n} s^i$ to a control action $u \in \mathcal{U}$, drawn from a space of controls $\mathcal{U} \subseteq \mathbb{R}^2$. We assume that the robot is not aware of agents' destinations s_T^i or policies $\pi_i : \mathcal{W}^{n+1} \times \mathcal{U}^i \rightarrow \mathcal{U}^i$, $i \in \mathcal{N}$. In this paper, our goal is to design a policy π that enables the robot to navigate from s_0 to s_T safely and socially.

4 Group-based Prediction

We introduce a framework for group-based representations based on [13] and a model for group-based prediction that is amenable for use in decentralized multiagent navigation.

4.1 Group Representation

Define as $\theta^i \in [0, 2\pi)$ the orientation of agent $i \in \mathcal{N}$ which is assumed to be aligned with the direction of its velocity u^i, extracted via finite differencing of its position over a timestep dt and denote by $v^i = ||u^i|| \in \mathbb{R}^+$ its speed. We define an augmented state for agent i as $q^i = (s^i, \theta^i, v^i)$.

We treat a social group as a set of agents who are in close proximity and share similar motion characteristics. Assume that a set of J groups, $\mathcal{J} = \{1, \ldots, J\}$ navigate in a scene. Define by $g^i \in \mathcal{J}$ a label indicating the group membership of agent i. We then define a group $j \in \mathcal{J}$ as a set $G^j = \{i \in \mathcal{N} \mid g^i = j\}$ and collect the set of all groups in a scene into a set $\boldsymbol{G} = \{G^j \mid j \in \mathcal{J}\}$.

Extracting Group Membership. We define the combined augmented state of all agents as $\boldsymbol{q} = \cup_{i=1:n} q^i$. To obtain group memberships for a set of agents $\mathcal{N}$, we apply the Density-Based Spatial Clustering of Applications with Noise algorithm (DBSCAN) [52] on agent states:

$$\boldsymbol{G} \longleftarrow \texttt{DBSCAN}(\boldsymbol{q} \mid \epsilon_s, \epsilon_\theta, \epsilon_v) \tag{1}$$

Where $\epsilon_s, \epsilon_\theta, \epsilon_v$ are respectively threshold values on agent distances, orientation and speeds.

Extracting the Social Group Space. For each group G^j, $j \in \mathcal{J}$, we define a *social group space* as a geometric enclosure $\mathcal{G}^j$ around agents of the group. For each agent $i \in G^j$, we define a personal space $\mathcal{P}^i$ as a two-dimensional asymmetric Gaussian based on the model introduced by Kirby [53]. Refer to Appendix A for detailed descriptions.

Given the personal spaces $\mathcal{P}^i$, $i \in G^j$, of all agents in a group j, we extract the social group space of the whole group as a convex hull:

$$\mathcal{G}^j = \text{Convexhull}(\{\mathcal{P}^i \mid i \in G^j\}). \tag{2}$$

The shape described by $\mathcal{G}^j$ represents an obstacle space representation of a group containing agents in close proximity with similar motion characteristics. For convenience, let us collect the spaces of all groups in a scene into a set $\boldsymbol{\mathcal{G}} = \{\mathcal{G}^j \mid j \in \mathcal{J}\}$.

4.2 Group Space Prediction Oracle

Based on the group-space representation of Sec. 4.1, we describe a prediction oracle that outputs an estimate of the future spaces occupied by a set of groups $\boldsymbol{\mathcal{G}}_{t:t_f}$ up to a time $t_f = t + f$, where f is a future horizon given a past sequence of group spaces $\boldsymbol{\mathcal{G}}_{t_h:t}$ from time $t_h = t - h$ where h is a window of past observations:

$$\boldsymbol{\mathcal{G}}_{t:t_f} \leftarrow \mathcal{O}(\boldsymbol{\mathcal{G}}_{t_h:t}) = \cup_{j=1:J} \mathcal{O}_j(\mathcal{G}^j_{t_h:t}), \tag{3}$$

Table 1: Autoencoder Performance

	Metric	ETH	HOTEL	ZARA1	ZARA2	UNIV
Baseline	mIoU (%)	83.52	90.37	88.04	89.30	85.32
	fIoU (%)	76.32	85.38	82.14	83.88	77.24
Autoencoder	mIoU (%)	86.66	92.10	89.97	90.94	87.52
	fIoU (%)	78.64	86.83	83.77	85.09	78.55

where $\mathcal{O}_j$ is a model generating a group space prediction for group G^j. Refer to Appendix B for detailed description of partial input handling.

We implement the oracle $\mathcal{O}_j$ of eq. (3) using a simple autoencoder. The encoder follows the 3D convolutional architecture in [54] whereas the decoder mirrors the model layout of the encoder. The autoencoder takes as input a sequence[1] $\mathcal{G}_{t_h:t}$ and outputs a sequence $\mathcal{G}_{t+1:t_f}$ which we pass through a sigmoid layer. We supervise the autoencoder's output using the binary cross entropy loss.

We verified the effectiveness of our autoencoder on the 5 scenes of our experiments by conducting a cross-validation comparison against a baseline. The baseline predicts the future shapes by linearly translating the last social group shape using its geometric center velocity. We use Intersection over Union (IoU) as our metric. Between the ground truths and the predictions, this metric divides the number of overlapped pixels by the number of pixels occupied by either one of them. As shown in Table 1, our autoencoder outperforms the baseline.

5 Model Predictive Control with Group-based Prediction

We describe G-MPC, a model predictive control (MPC) framework for navigation in multiagent environments that leverages the group-based prediction oracle of Sec. 4.

We describe our group-prediction informed MPC, or G-MPC. At planning time t, given a (possibly partial) augmented world state history $\boldsymbol{Q}_{t_{\hat{h}}:t}$, we first extract a sequence of group spaces $\boldsymbol{\mathcal{G}}_{t_h:t}$ based on the method of Sec. 4.1. Given these, the robot computes an optimal control trajectory $\boldsymbol{u}^* = u^*_{1:K}$ of length K by solving the following optimization problem:

$$(\boldsymbol{s}^*, \boldsymbol{u}^*) = \arg\min_{u_{1:K}} \sum_{k=1:}^{K} \gamma^k J(s_{k+1}, \boldsymbol{\mathcal{G}}_{k+1}, s_T) \tag{4}$$

$$s.t.\ \boldsymbol{\mathcal{G}}_{2-h:1} \leftarrow \boldsymbol{\mathcal{G}}_{t_h:t} \tag{5}$$

$$s_1 \leftarrow s_t \tag{6}$$

$$\boldsymbol{\mathcal{G}}_{k+1:k_f} = \mathcal{O}(\boldsymbol{\mathcal{G}}_{k_h:k}) \tag{7}$$

$$u_k \in \mathcal{U} \tag{8}$$

$$s_{k+1} = s_k + u_k \cdot dt, \tag{9}$$

where γ is the discount factor and J represents a cost function, eq. (5) initializes the group space history ($k = 2 - h$ is the timestep displaced a horizon h in the past from the first MPC-internal timestep $k = 1$), eq. (6) initializes the robot state to the current robot state s_t, eq. (7) is an update rule recursively generating a predicted future group sequence up to timestep $k_f = k + f$ given history from time $k_h = k - h$ up to time k, $\mathcal{O}$ represents a group-space prediction oracle based on Sec. 4, and eq. (9) is the robot state transition assuming a fixed time parametrization of step size dt.

We employ a weighted sum of costs J_g and J_d, penalizing respectively distance to the robot's goal and proximity to groups:

$$J(s_k, \boldsymbol{\mathcal{G}}_k, s_T) = \lambda J_g(s_k, s_T) + (1 - \lambda) J_d(s_k, \boldsymbol{\mathcal{G}}_k), \tag{10}$$

where λ is a weight representing the balance between the two costs and

$$J_g(s_k) = \begin{cases} 0, & \text{if } s_k \in \boldsymbol{\mathcal{G}}_k \\ ||s_{k-1} - s_T||, & \text{else,} \end{cases} \tag{11}$$

[1]The oracle input sequence is first converted into image-space coordinates using the homography matrix of the scene. We also preprocess inputs to have normalized scale and group positions. The autoencoder output is converted back into Cartesian coordinates using the inverse homography transform.

Table 2: Number of trials per task and scene.

Task	ETH	HOTEL	ZARA1	ZARA2	UNIV
Flow	58	43	25	127	106
Cross	58	44	28	129	114

penalizes a rollout according to the distance of the last collision-free waypoint to the robot's goal. Further, we define J_d as:

$$J_d(s_k, \boldsymbol{\mathcal{G}}_k) = \exp(-\mathcal{D}\left(s_{k+1}, \boldsymbol{\mathcal{G}}_k\right)), \tag{12}$$

where

$$\mathcal{D}(s_k, \boldsymbol{\mathcal{G}}_k) = \begin{cases} \min_{j \in \mathcal{J}} D\left(s_k - \mathcal{G}_k^j\right), & \text{if } s_k \notin \mathcal{G}_k^j \\ -\min_{j \in \mathcal{J}} D\left(s_k - \mathcal{G}_k^j\right), & \text{else,} \end{cases} \tag{13}$$

where $D(s_k - \mathcal{G}_k^j)$ returns the minimum distance between the robot state and the space occupied by group j at time k. Using D, function $\mathcal{D}$ computes the minimum distance to any group for a given time. In most cases, the robot lies outside of groups, i.e., $s_k \notin \mathcal{G}_k^j$ –therefore, the cost J_d tries to maximize the distance $\mathcal{D}$. Sometimes, the robot might end up entering the group space $\boldsymbol{\mathcal{G}}$ –in those cases, J_d tries to minimize $\mathcal{D}$, to steer the robot towards the direction of quickest escape from the group. In case that the robot is inside a group to begin with, we shrink the group sizes in Sec. 4.1 until the robot is outside the groups again.

To solve eq. (4), we search over a finite set $\boldsymbol{\mathcal{U}}$ of control trajectories of horizon K. With the assumption that the robot is holonomic and is not under any kinematic constraints, we use a set of R control rollouts $\boldsymbol{\mathcal{U}} = \{\boldsymbol{u}^1, ..., \boldsymbol{u}^R\}$ with three levels of tangential speeds and a set of turning speed, i.e.,

$$u_{1:K}^r = (v\cos\psi, v\sin\psi, \omega),\ \psi = \frac{2\pi r}{R}, v \in \left\{\frac{1}{3}v_{max}, \frac{2}{3}v_{max}, v_{max}\right\}, \omega \in \left\{0, \pm\frac{\pi}{2}\right\} \tag{14}$$

To ensure compatibility between our group-based prediction model and our MPC formulation, we set the control rollout time horizon to be the prediction model's prediction horizon, or $K = f$.

6 Evaluation

We evaluate our framework through a simulation study in which the robot performs a navigation task (a transition between two points) within a crowds of dynamic agents in a set of scenes.

6.1 Experimental Setup

We consider a set of realistic pedestrian scenes, drawn from the ETH [14] (ETH and HOTEL scenes) and UCY [15] (ZARA1, ZARA2 and UNIVERSITY scenes) datasets, which often serve as benchmarking testbeds in the motion prediction and social navigation literature [55, 17, 56, 57]. In each scene, we define two navigation tasks (see Fig. 2): *Flow*: in which the robot navigates along the crowd flow and *Cross* in which the robot intersects vertically with the traffic flow. For each task, we generate a set of trials by segmenting the scene recording into blocks involving challenging interactions. We define a challenging interaction to be a segment involving at least 5 pedestrians inside the test region drawn in black in Fig. 2. This process provided us with a distribution of trials as shown in table Table 2. Across all trials, we keep the robot's maximum speed at $1.75m/s$.

We consider two experimental conditions: an *Offline* and an *Online* one. In the *Offline* one, the robot navigates among a crowd moving according to a recording of a human crowd. Under this condition, pedestrians act as dynamic obstacles that do not react to the robot, a situation which could arise in cases where robots are of shorter size and could thus be easily missed by navigating pedestrians. In the *Online* one, the robot navigates among a crowd[2] moving by running ORCA [16], a policy that is frequently used as a simulation engine for benchmarking in the social navigation literature[57, 8, 58].

To investigate the value of G-MPC, we develop three variants of it. **group-auto** is a G-MPC in which the autoencoder has a history $h = 8$ and a horizon $f = 8$. **group-nopred** is a variant that

[2]For consistency, the agents in the crowd start and end at the same spots as the agents in the recorded crowd from the Offline condition.

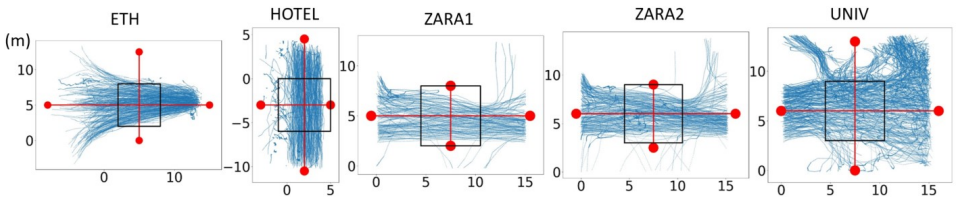

Figure 2: Trajectories of all pedestrians in the datasets. The red dots represent the task start and end locations. The red lines represent the task paths. The black box represents the test region to check for non-trivial tasks.

features no prediction at all –it just reacts to observed groups at every timesteps and it is equivalent to the framework of Yang and Peters [26]. Finally, **laser-group-auto** is identical to **group-auto** but instead of using ground-truth pose information, it takes as input noisy lidar scan readings. We simulate this by modeling pedestrians as $1m$-diameter circles and lidar scans as rays projecting from the robot. We refer to the spec sheet of a SICK LMS511 2D lidar for simulation parameters. We further inject noise into the readings according to the spec sheet. Under this simulation, pedestrians may only be partially observable or even completely occluded from the robot.

We compare the performance of these policies against a set of MPC variants using mechanisms for individual motion prediction. **ped-nopred** is a vanilla MPC that reacts to the current states of other agents without making predictions about their future states. **ped-linear** is a vanilla MPC that estimates future states of agents by propagating agents' current velocities forward. This baseline is motivated by recent work showing that constant-velocity models yield competitive performance in pedestrian motion prediction tasks [59]. Finally, **ped-sgan** is an MPC that uses S-GAN [17] to extract a sequence of future state predictions for agents based on inputs of their past states. We selected S-GAN because it is a recent highly performing model. To ensure a fair comparison, all the MPC policy variants are integrated with the same MPC controller evaluated at $dt = 0.1$.

We measure the performance of the policies with respect to four different metrics: a) *Success rate*, defined as the ratio of successful trials over total number of trials; b) *Comfort*, defined as the ratio of trials in which the robot does not enter any social group space over the total number of trials; c) *Minimum distance to pedestrians*, defined as the smallest distance between the robot and any agent per trial; d) *Path length*, defined as the total distance traversed by the robot in a trial.

To track the performance of G-MPC, we design a set of hypotheses targeting aspects of safety and group space violation which we investigate under both experimental conditions, i.e., offline and online:

H1: To explore the benefits of group based representations alone, we hypothesize that **group-nopred** is safer than **ped-nopred** while achieving similar success rates but worse efficiency.

H2: To explore the full benefit of group based formulation, we hypothesize that **group-auto** is safer than **ped-linear** and **ped-sgan** while achieving similar success rates but worse efficiency.

H3: To explore how our formulation handles imperfect inputs, we hypothesize that **laser-group-auto** achieves similar safety to **group-auto** while achieving similar success rate and efficiency.

H4: To check that our formulation is socially compliant, we hypothesize that **group-nopred**, **group-auto** and **laser-group-auto** violate agents' group space less often than the baselines.

6.2 Results

Quantitative Analysis. Fig. 3 and Fig. 4 contain bar charts representing the performance of G-MPC compared with its baselines under Offline and Online settings respectively. Bars indicate means, errorbars indicate standard deviations, "F" and "C" are flow and cross scenarios respectively, and the number of asterisks indicates increasing significance levels: $\alpha = 0.05, 0.01, 0.001$ according to two-sided Mann-Whitney U-tests.

H1: We can see from both Fig. 3 and Fig. 4 that G-MPC achieves statistically significantly larger minimum distances to pedestrians across all scenarios, often with $p < 0.001$. This illustrates that the group representation is in itself capable of upgrading a simple MPC with no prediction. As expected, we observe that the price G-MPC pays for that is a larger average path length. We also see that success rates are comparable. Overall, we conclude that H1 holds.

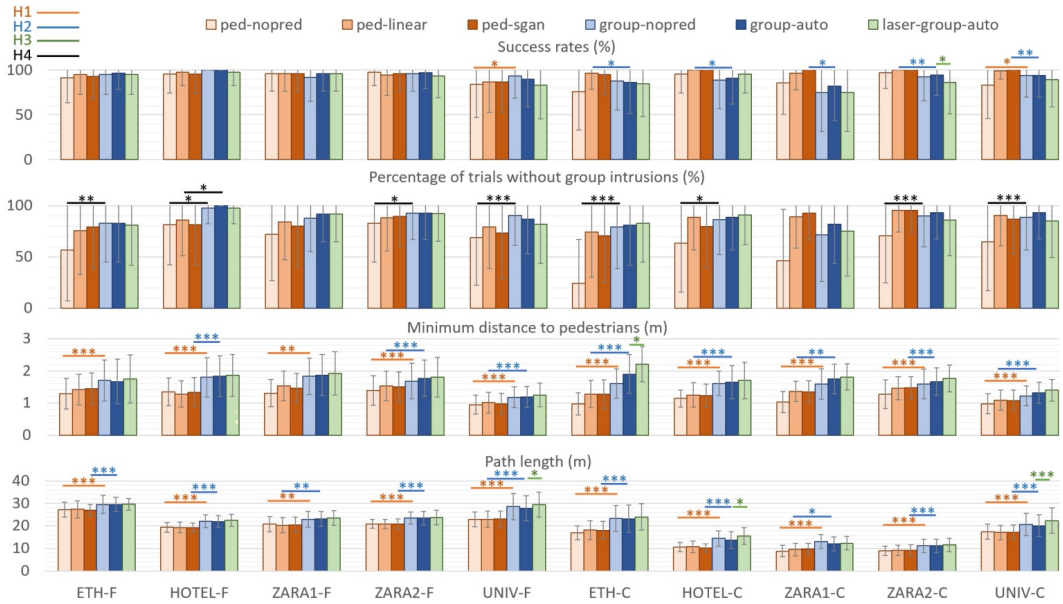

Figure 3: Performance per scene under the *Offline* condition. Horizontal lines indicate statistically significant results corresponding to different to hypotheses.

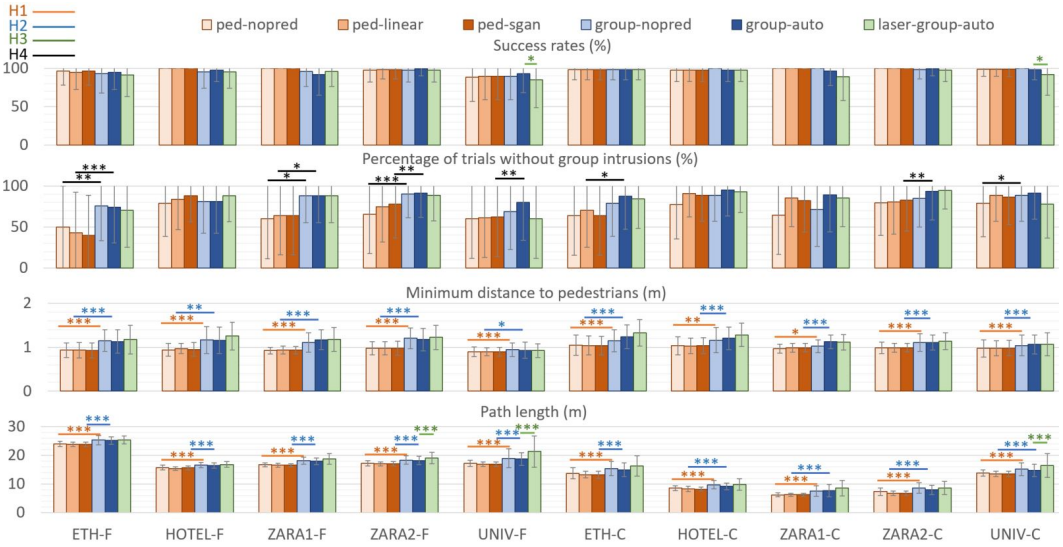

Figure 4: Performance per scene under the *Online* condition (simulated pedestrians powered by ORCA [16]). Horizontal lines indicate statistically significant results corresponding to different to hypotheses.

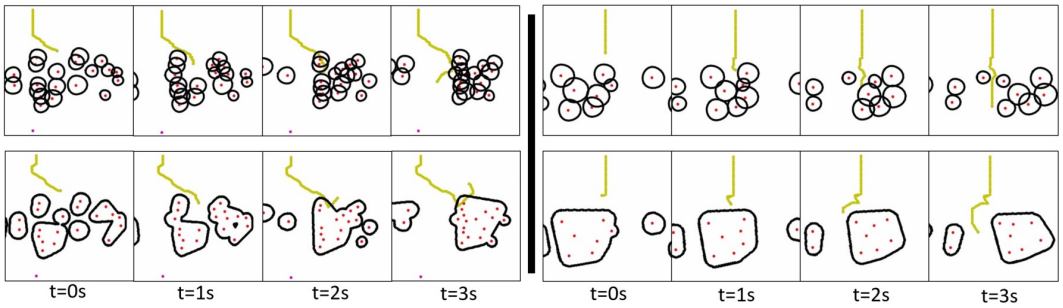

Figure 5: Qualitative performance difference between approaches leveraging pedestrian-based (top) and group-based (bottom) representations. Left: non-reactive agents. Right: reactive agents. In all cases, the robot starts from the top and attempts to navigate downward. The green line shows the robot's traversed trajectory.

H2: When future state predictions are considered, G-MPC obtains statistically significant results in most scenes supporting its attributes of being safer at the cost of worse efficiency. Thus H2 is partially confirmed. In offline scenarios, G-MPC has lower success rates in crossing scenarios. Upon closer inspection, most failure cases are due to timeouts from G-MPC's conservative behavior. However, in online scenarios where pedestrians react to the robot, G-MPC achieves high success rates. In real-world situations, to cross dense traffic, the robot needs to plan its actions with expectations of reactive pedestrians. Otherwise, the robot will most probably run into *the freezing robot problem* [4].

H3: Group-based representations have the potential to robustly account for imperfect state-estimates. Overall, we observe that with simulated imperfect states, G-MPC does not perform statistically significantly worse in terms of safety, but in dense crowds of the UNIV scenes it has worse efficiency and worse success rates in online cases. This shows that H3 holds in terms of safety and, in moderately dense human crowds, holds in terms of efficiency. Future work on better group representation is needed to achieve better efficiency in high-density human crowds given imperfect states.

H4: From Fig. 3 and Fig. 4, we can see that G-MPC often has fewer group-space intrusions than its baselines. While this relationship is not always statistically significant, we do see a general trend of the group-based approaches to respect group spaces more often than individual ones. Thus, we conclude that H4 is partially confirmed.

We additionally observe a general trend that **group-auto** is better than **group-nopred** in terms of higher success rates, lower chances of group intrusions, longer minimum distances to pedestrians and shorter path lengths. This shows that our group prediction model offers benefits to the robot's navigation. However, in a few scenarios **group-nopred** performs better. We largely attribute this to the finite inaccuracies of future group predictions and the freezing robot problem that accompanies the robot's more conservative behavior in **group-auto** than in **group-nopred**.

Qualitative Analysis. Qualitatively, it is a more common occurrence for regular MPC to perform aggressive and socially inappropriate maneuvers than G-MPC. As shown in the two examples in Fig. 5 executed by **ped-sgan** and **group-auto** agents, we can see that in offline conditions, the MPC agent aggressively cuts in front of the two pedestrians to the left before proceeding headlong into the cluster of pedestrians, only managing to avoid the deadlock by escaping through the narrow gap that opens up. While for G-MPC, it tracks the movements of the two pedestrian groups coming from the left. When the two pedestrian groups merge, the agent turns around and reevaluates its approach to cross. In the online condition, we observe that the MPC agent cuts through a pedestrian group to reach the other side, forcing a member of the group to stop and yield as indicated by the pedestrian's shrinking personal space, which is proportional to its speed. In the same situation, the G-MPC agent chooses to circumvent behind the social group.

7 Conclusion

We introduced a methodology of generating group-based representations and predicting their future states. Through an extensive evaluation over the flow and crossing scenarios drawn from 10 different real-world scenes from 2 different human datasets with both reactive and non-reactive agents, we demonstrate that our approach is safer and more socially compliant. Through experimentation with simulated laser scans, our model displays promising potential to process noisy sensor inputs without much performance downgrade.

Various improvements to our control framework are possible. For example, we could incorporate state-of-the-art oracles in the form of advanced video prediction models [60] or incorporate inter-group interaction modeling. Additional considerations such as the set of rollouts or the cost functions could possibly increase performance. Finally, alternative control frameworks such as reinforcement learning approaches could be applicable. However, our goal in this paper was to illustrate the value of group-based representations for navigation tasks. Future work will involve improving both the prediction and the control components of our framework.

Finally, we plan on validating our findings on a real-world robot to fully test the capability of G-MPC to handle noisy sensor inputs. We also plan to investigate ways to improve computation time to enhance our approach's real-world applicability, which includes simplifying group representation geometry and predicting future group states in metric space instead of in image space.

References

[1] S. Thrun, M. Bennewitz, W. Burgard, A. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz. MINERVA: a second-generation museum tour-guide robot. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, volume 3, pages 1999–2005, 1999.

[2] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch. Human-aware robot navigation: A survey. *Robotics and Autonomous Systems*, 61(12):1726–1743, 2013.

[3] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard. Socially compliant mobile robot navigation via inverse reinforcement learning. *The International Journal of Robotics Research*, 35 (11):1289–1307, 2016.

[4] P. Trautman, J. Ma, R. M. Murray, and A. Krause. Robot navigation in dense human crowds: Statistical models and experimental studies of human-robot cooperation. *International Journal of Robotics Research*, 34(3):335–356, 2015.

[5] C. I. Mavrogiannis, A. M. Hutchinson, J. Macdonald, P. Alves-Oliveira, and R. A. Knepper. Effects of distinct robotic navigation strategies on human behavior in a crowded environment. In *Proceedings of the 2019 ACM/IEEE International Conference on Human-Robot Interaction (HRI '19)*. ACM, 2019.

[6] M. Luber, L. Spinello, J. Silva, and K. Arras. Socially-aware robot navigation: A learning approach. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 902–907, 2012.

[7] B. Kim and J. Pineau. Socially adaptive path planning in human environments using inverse reinforcement learning. *International Journal of Social Robotics*, 8(1):51–66, 2016.

[8] M. Everett, Y. F. Chen, and J. P. How. Motion planning among dynamic, decision-making agents with deep reinforcement learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, Sept. 2018.

[9] C. Mavrogiannis, V. Blukis, and R. A. Knepper. Socially competent navigation planning by deep learning of multi-agent path topologies. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6817–6824, 2017.

[10] K. Koffka. *Principles of Gestalt psychology*. Harcourt, Brace, 1935.

[11] A. Desolneux, L. Moisan, and J.-M. Morel. *From Gestalt Theory to Image Analysis: A Probabilistic Approach*. Springer Publishing Company, Incorporated, 1st edition, 2007. ISBN 0387726357.

[12] M. Vázquez and A. Steinfeld. An assisted photography method for street scenes. In *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 89–94, 2011.

[13] A. Wang and A. Steinfeld. Group split and merge prediction with 3D convolutional networks. *IEEE Robotics and Automation Letters*, 5(2):1923–1930, 2020.

[14] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 261–268, Sept 2009.

[15] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. *Comput. Graph. Forum*, 26(3):655–664, 2007.

[16] J. van den Berg, S. J. Guy, M. Lin, and D. Manocha. Reciprocal n-body collision avoidance. In *Robotics Research*, pages 3–19. Springer Berlin Heidelberg, 2011.

[17] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social GAN: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2255–2264, 2018.

[18] C. I. Mavrogiannis, W. B. Thomason, and R. A. Knepper. Social momentum: A framework for legible navigation in dynamic multi-agent environments. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*, pages 361–369. ACM, 2018.

[19] C. Chen, Y. Liu, S. Kreiss, and A. Alahi. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 6015–6022, 2019.

[20] W. Zhi, T. Lai, L. Ott, and F. Ramos. Anticipatory navigation in crowds by probabilistic prediction of pedestrian future movements. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[21] A. Somani, N. Ye, D. Hsu, and W. S. Lee. Despot: Online pomdp planning with regularization. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[22] P. Cai, Y. Luo, D. Hsu, and W. S. Lee. Hyp-despot: A hybrid parallel algorithm for online planning under uncertainty. *The International Journal of Robotics Research*, 40(2-3):558–573, 2021.

[23] S. Kiss, K. Katuwandeniya, A. Alempijevic, and T. A. Vidal-Calleja. Anticipatory navigation in crowds by probabilistic prediction of pedestrian future movements. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[24] J. Šochman and D. C. Hogg. Who knows who - inverting the social force model for finding groups. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 830–837, 2011.

[25] A. Kendon. Conducting interaction : Patterns of behavior in focused encounters. *Studies in International Sociolinguistics*, 7, 1990.

[26] F. Yang and C. Peters. Social-aware navigation in crowds with static and dynamic groups. In *2019 11th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pages 1–4, 2019.

[27] L. Bazzani, M. Cristani, and V. Murino. Decentralized particle filter for joint individual-group tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1886–1893, 2012.

[28] M. Chang, N. Krahnstoever, and W. Ge. Probabilistic group-level motion analysis and scenario recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 747–754, 2011.

[29] G. Gennari and G. D. Hager. Probabilistic data association methods in visual tracking of groups. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–II, 2004.

[30] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010*, pages 452–465, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15549-9.

[31] M. Zanotto, L. Bazzani, M. Cristani, and V. Murino. Online bayesian nonparametrics for group detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 111.1–111.12, 2012.

[32] I. Chamveha, Y. Sugano, Y. Sato, and A. Sugimoto. Social group discovery from surveillance videos: A data-driven approach with attention-based cues. In *Proceedings of the The British Machine Vision Association (BMVC)*, 2013.

[33] S. D. Khan, G. Vizzari, S. Bandini, and S. Basalamah. Detection of social groups in pedestrian crowds using computer vision. In S. Battiato, J. Blanc-Talon, G. Gallo, W. Philips, D. Popescu, and P. Scheunders, editors, *Advanced Concepts for Intelligent Vision Systems*, pages 249–260. Springer International Publishing, Cham, 2015.

[34] D. Helbing and P. Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51 (5):4282–4286, 1995.

[35] R. Mazzon, F. Poiesi, and A. Cavallaro. Detection and tracking of groups in crowd. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 202–207, 2013.

[36] F. Solera, S. Calderara, and R. Cucchiara. Socially constrained structural learning for groups detection in crowd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5): 995–1008, 2016.

[37] W. Ge, R. T. Collins, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):1003–1016, 2012.

[38] A. Taylor, D. M. Chan, and L. D. Riek. Robot-centric perception of human groups. *ACM Transactions on Human-Robot Interaction*, 9(3):1–21, 2020.

[39] I. Chatterjee and A. Steinfeld. Performance of a low-cost, human-inspired perception approach for dense moving crowd navigation. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 578–585, Aug 2016.

[40] N. P. Cuntoor, R. Collins, and A. J. Hoogs. Human-robot teamwork using activity recognition and human instruction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 459–465, 2012.

[41] R. Gockley, J. Forlizzi, and R. Simmons. Natural person-following behavior for social robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 17–24, 2007.

[42] C. Granata and P. Bidaud. A framework for the design of person following behaviors for social mobile robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4652–4659, 2012.

[43] E. Jung, B. Yi, and S. Yuta. Control algorithms for a mobile robot tracking a human in front. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2411–2416, 2012.

[44] H. Zender, P. Jensfelt, and G. M. Kruijff. Human- and situation-aware people following. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1131–1136, 2007.

[45] A. Nanavati, X. Z. Tan, J. Connolly, and A. Steinfeld. Follow the robot: Modeling coupled human-robot dyads during navigation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3836–3843, 2019.

[46] D. Feil-Seifer and M. Matarić. People-aware navigation for goal-oriented behavior involving a human partner. In *Proceedings of the IEEE International Conference on Development and Learning (ICDL)*, volume 2, pages 1–6, 2011.

[47] A. K. Pandey and R. Alami. A step towards a sociable robot guide which monitors and adapts to the person's activities. In *2009 International Conference on Advanced Robotics*, pages 1–8, 2009.

[48] A. Garrell and A. Sanfeliu. Local optimization of cooperative robot movements for guiding and regrouping people in a guiding mission. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3294–3299, 2010.

[49] M. Shiomi, T. Kanda, S. Koizumi, H. Ishiguro, and N. Hagita. Group attention control for communication robots with wizard of oz approach. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 121–128, 2007.

[50] E. A. Martinez-Garcia, Ohya Akihisa, and Shin'ichi Yuta. Crowding and guiding groups of humans by teams of mobile robots. In *Proceedings of the IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, pages 91–96, 2005.

[51] K. Katyal, Y. Gao, J. Markowitz, I.-J. Wang, and C.-M. Huang. Group-Aware Robot Navigation in Crowded Environments. *arXiv e-prints*, art. arXiv:2012.12291, Dec. 2020.

[52] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. Int. Conf. Knowl. Discovery and Data Mining*, pages 226–231, 1996.

[53] R. Kirby. *Social Robot Navigation*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, May 2010.

[54] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, December 2015.

[55] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016.

[56] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 12085–12094, June 2019.

[57] C. Cao, P. Trautman, and S. Iba. Dynamic channel: A planning framework for crowd navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5551–5557, 2019.

[58] C. Mavrogiannis, F. Baldini, A. Wang, D. Zhao, P. Trautman, A. Steinfeld, and J. Oh. Core Challenges of Social Robot Navigation: A Survey. *arXiv e-prints*, art. arXiv:2103.05668, Mar. 2021.

[59] C. Schöller, V. Aravantinos, F. Lay, and A. Knoll. What the constant velocity model can teach us about pedestrian motion prediction. *IEEE Robotics and Automation Letters*, 5(2):1696–1703, 2020.

[60] V. L. Guen and N. Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.