Forecasting Interactive Dynamics of Pedestrians with Fictitious Play

Wei-Chiu Ma[1] De-An Huang[2] Namhoon Lee[3] Kris M. Kitani[4]

[1]MIT [2]Stanford [3]Oxford [4]CMU

Abstract

*We develop predictive models of pedestrian dynamics by encoding the coupled nature of multi-pedestrian interaction using game theory and deep learning-based visual analysis to estimate person-specific behavior parameters. We focus on predictive models since they are important for developing interactive autonomous systems (*e.g.*, autonomous cars, home robots, smart homes) that can understand different human behavior and pre-emptively respond to future human actions. Building predictive models for multi-pedestrian interactions however, is very challenging due to two reasons: (1) the dynamics of interaction are complex interdependent processes, where the decision of one person can affect others; and (2) dynamics are variable, where each person may behave differently (*e.g.*, an older person may walk slowly while the younger person may walk faster). We address these challenges by utilizing concepts from game theory to model the intertwined decision making process of multiple pedestrians and use visual classifiers to learn a mapping from pedestrian appearance to behavior parameters. We evaluate our proposed model on several public multiple pedestrian interaction video datasets. Results show that our strategic planning model predicts and explains human interactions 25% better when compared to a state-of-the-art activity forecasting method.*

1. Introduction

The goal of this work is to imitate the predictive abilities of human cognition, by building a predictive model that takes into account complex reasoning about: (1) the interdependent interactions of multiple pedestrians and (2) important visual cues needed to infer individual behavior patterns. Consider the complexities of predicting the trajectories of multiple pedestrians from a *single* image, as depicted in Figure 1 where four pedestrians are walking on the street. Given this single image, what would one forecast as their future trajectories? A simple prediction would be that all people will walk in a straight line (*i.e.*, the minimum distance) to their goal as in Figure 1(b). This strategy, however, might lead to collisions between pedestrians

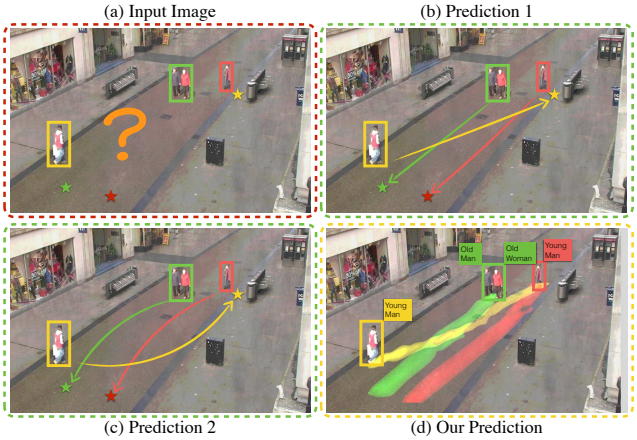

Figure 1. Can you forecast their future behavior? A single image contains rich information about future trajectories.

(*e.g.*, the young man (yellow) and the older couple (green) may collide). A more thoughtful model might consider the possibility that one or more pedestrians will alter their trajectory based on their *prediction of other pedestrians* (Figure 1(c)). Going further, a more informed model might attempt to take into account the observation that on average, elderly couples tend to walk at a slower rate, while a young man is more likely walk quickly and take pre-emptive maneuvers (Figure 1(d)). Taking these observations into consideration, we might hypothesize that the younger man is more likely to exemplify preemptive avoidance behaviors and weave through the two pedestrians. This illustration serves to highlight the complex reasoning that is involved in predicting the walking trajectories of several people given limited amount of information (in our scenario a single image). Our goal is to mimic – computationally – this ability to reason about the dynamics of interactive social processes.

Developing computational models for interactive dynamics among humans, however, is an extremely challenging task. This requires a deep understanding of the complex and often subtle norms of human interactions. Pioneering works have attempted to address this by parameterizing human behaviors with models such as social forces [13, 30], potential fields [2], and flow fields [3, 4]. Yet most of these works are performing either long-term prediction in static

environments or short-term prediction in dynamic environments. They do not address the interactions that could occur in the distant future, and do not resolve the long-term prediction problem in dynamic environments. To address these complexities of multi-agent future prediction, we propose a game-theoretic approach.

We directly address the interdependent nature of human interactions using the language and concepts of multi-player game theory. In particular, we utilize Brown's [8] classical notion of *Fictitious Play* to model the interaction between multiple pedestrians. Brown's fictitious play model assumes that each player will take the best next action based on an observed empirical distribution over the past strategies of other players. As we will show, the multi-player game model has strong parallels to multi-pedestrian forecasting, as each pedestrian pre-emptively plans her path according to beliefs about how other pedestrians will move.

To individualize the pedestrian model, we train a deep learning-based classifier to learn visual cues that are indicative of behavior patterns (*e.g.*, age can affect speed). We use the classifier to estimate each pedestrian's velocity based on sub-population statistics. Furthermore, we visually estimate the initial body orientation such that the model is more likely to predict motion aligned to body direction at the start of a predicted trajectory. In this way, we integrate visual analysis with our prediction model. Figure 2 shows the overview of our approach.

Contributions: We present a novel technique to forecast multi-pedestrian trajectories from a *single* image. First, we explicitly model the interplay among multiple people by drawing connections between game theory and optimal control. To the best of our knowledge, Fictitious Play has never been applied in the context of modeling pedestrian motions. Second, we address the variability among people by building individualized predictive pedestrian models. We are the first attempt to infer physical properties of each pedestrian from appearance for multi-agent forecasting.

2. Related Work

There has been growing interest in developing computational models of human activities that can extrapolate unseen information and predict future unobserved activities [34, 38, 22, 33, 18, 42, 40, 37, 15, 14, 16, 17, 35, 9, 49, 43, 36, 32]. In the context of pedestrian dynamics, Helbing and Molonar [13] first integrated the concept of the *social force* model into a computational framework for understanding pedestrian dynamics. Their work incorporated ideas of goals, desired speed and the repulsion due to territorial affects of social forces. In computer vision, the social force model has been used to help aid visual tracking [30] and anomaly detection [26]. More recent work has focused on discovering the underlying potential field by observing human behavior such as patterns of motions [2], mutual

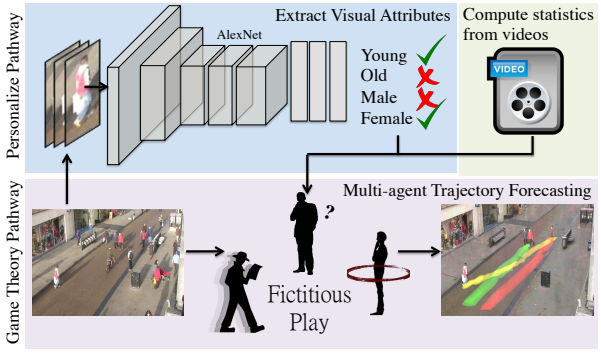

Figure 2. Model Overview. **Personalization pathway** (**top**) estimates physical properties for each pedestrian based on visual information and statistics from videos. **Game theory pathway** (**bottom**) takes as input: (1) estimated properties, (2) individualized motion model per pedestrian. As output, it forecasts multi-pedestrian interactions/trajectories using Fictitious Play.

gaze or regions of repulsion. In high-density crowds, the patterns of motion of people can be used to infer an underlying flow field for a given scene [3, 4] and the interaction between stationary crowds and pedestrians can be used to predict pedestrians' future motions [46]. The global motion or the joint attention of sparse groups of people (e.g., sports scenarios) can also be used to infer basins of attractions or socially salient hot spots [19, 29]. Patterns of avoidance can also be used to learn the hidden rewards or costs of physical spaces [21, 44, 41].

To make reliable predictions about the long-term future, many techniques often assume a static environment [21, 41]. In a static environment, the cost topology is constant, where the environment and features do not change over time. In dynamic environments, the cost topology of the state space is constantly changing which means that any computational model must be continually updated. When the cost topology can be accurately updated over time, it can be used for short-term prediction [30, 11, 20] (or at least until the next update). As such, these techniques have been very effective for tracking multi-pedestrian trajectories. While methods have been proposed for long-term prediction in static environments and short-term prediction in dynamic environments, the task of long-term prediction in dynamics environments remains relatively unexplored in human activity analysis except [1, 24, 18]. Concurrent with our work, [1] introduced a data-driven approach to *implicitly* encode the interactive dynamics among people. Their model, however, focused only on trajectory data. They ignored the rich information underlying the visual data. In [24, 18], the complex and intertwined interactions between agents is either ignored [18] or restricted to the perspective of a single agent (only the wide receiver in [24]). In constrast, we directly address the interdependent nature of human interactions using Fictitious Play and perform long-term prediction for *all* of the agents in the scene.

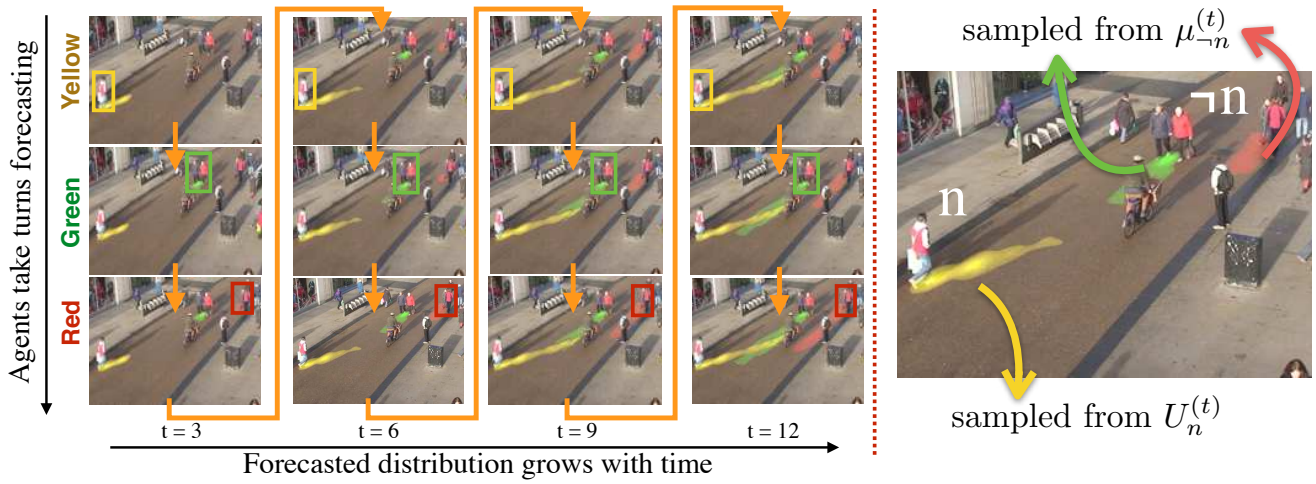

Figure 3. Left: Visualization of Fictitious Play with three pedestrians. Right: Distributions over states sampled from $U_n^{(t)}$ and $\mu_{\neg n}^{(t)}$.

3. Forecasting Multi-Pedestrian Trajectories

Given a single image and initial pedestrians detections, we aim to develop a predictive model that can forecast plausible future trajectories for all pedestrians. To do this, we must model the complex predictive interplay between multiple pedestrians, while also considering individual differences that might impact behavior, to obtain accurate predictions. To address these challenges, we utilize concepts from game theory to model the intricately coupled interactive prediction process. We also leverage recent success of deep neural networks to infer individual behavior models for each pedestrian from visual evidence. We describe how game theory can be used to frame our multi-pedestrian forecasting problem in Section 3.1 and present a method for mapping the visual appearance of pedestrians to estimate person-specific behavior parameters in Section 3.4.

Notation. We will define the state space (the ground plane) as a 2D lattice, where each position is denoted by $\boldsymbol{x} = [x, y] \in \boldsymbol{X}$. A pedestrian can make a transition from state to state by taking an action $\boldsymbol{a} \in A$ which in the case of a 2D lattice (grid world) is the velocity $[\dot{x}, \dot{y}]$. A trajectory is a sequence of state-action pairs, $\boldsymbol{s} = \{(\boldsymbol{x}_1, \boldsymbol{a}_1), \ldots, (\boldsymbol{x}_K, \boldsymbol{a}_K)\}$. Each state $\boldsymbol{x}$ has an associated vector of features $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}) \ldots f_J(\boldsymbol{x})]$, where $f_j(\boldsymbol{x})$ represent properties of that state such as the output of a visual classifier, the distance to an object or predicted presence of another pedestrian.

3.1. Forecasting Interactions as Fictitious Play

Game theory [27] is a widely applicable discipline that aims to model adversarial and collaborative interactions between *rational* decision-makers. It has been applied to a range of disciplines including economic theory [31], politics [7] and computer science [28]. More importantly, it is well-suited for modeling our multi-pedestrian prediction scenario, as the social dynamics of collision avoidance can be modeled as a collaborative multi-player game. To forecast long-term trajectories of multiple pedestrians, we utilize Fictitious Play (FP) [8], where we model each pedestrian to take a path based on her own predictions of how other pedestrians will move. By incrementally forward simulating pedestrian paths with this model, we can obtain a distribution over possible future paths over multiple people.

Formally, each pedestrian $n \in \{1, \ldots, N\}$ has the ability to choose a macro-action $\boldsymbol{s}_n \in S$ from a set of macro-actions. In our scenario, a macro-action $\boldsymbol{s}_n$ is a very short trajectory whose length L_n depends on the speed of the pedestrian n (detailed in Section 3.4). Each pedestrian has an utility function $U_n[\boldsymbol{s}_n, \mu_{\neg n}(\boldsymbol{s}_{\neg n})]$ that maps a given macro-action to a value $U_n : \boldsymbol{s}_n \rightarrow \mathbb{R}$. Intuitively, the utility function U_n describes the reward of taking a certain path. If there is a low potential of collision, its utility will be high. Notice that the U_n is also dependent on the forecasted distributions over macro-actions of all other pedestrians $\mu_{\neg n}(\boldsymbol{s}_{\neg n})$. This is needed to compute the potential of collision with other pedestrians. The set of trajectories $\boldsymbol{s}_{\neg n}$ is a set of macro-actions of all other pedestrians, $\boldsymbol{s}_{\neg n} = \{\boldsymbol{s}_m | m \neq n\}$. We visualize a distribution over states sampled from $\mu_{\neg n}$ in Figure 3 (right).

Algorithm 1 describes the process of Fictitious Play. For every forecasting period τ, each pedestrian n forms beliefs about the future actions of other pedestrians by updating the empirical distribution $\mu_{\neg n}^{(t)}$ using the function UPDATEEMPIRICAL. Then the distribution $\mu_{\neg n}^{(t)}$ is encoded as social feature $f_{n,soc}^{(t)}$ using the function ENCODETOFEATURE. The utility function of the n-th pedestrians is updated according to this new feature with the function UPDATEUTILITY. In the final step, we forecast the movement of the pedestrian with the function TAKEMACROACTION.

Algorithm 2: EncodeToFeature

Input : Empirical distribution $\mu_{\neg n}^{(t)}$, State visitation distribution $D_{\neg n}^{(t-1)}$
Output: Feature vector $f_{n,soc}^{(t)}$

$f_{n,soc}^{(t)} = \mathbf{0}$
for $m = 1 : N$ *and* $m \neq n$ **do**
  $D_m^{(t:t+L_m)} \leftarrow$ TakeMacroAction$(\mu_m^{(t)}, D_m^{(t-1)})$
  $\bar{D}_m = \sum_{l=t}^{t+L_m} D_m^{(l)}$
  $f_{n,soc}^{(t)} = f_{n,soc}^{(t)} + \bar{D}_m$
end

Algorithm 3: TakeMacroAction

Input : Empirical distribution $\mu_n^{(t)}$, Prior state visitation distribution $D_n^{(t-1)}$
Output: Future state visitation distributions $D_n^{(t:t+L_n)}$

$\pi(\boldsymbol{a}|\boldsymbol{x}) \leftarrow$ ComputePolicy(μ_n)
for $l = t : t + L_n$ **do**
  $D_n^{(l)}(\boldsymbol{x}') = \sum_{\boldsymbol{a},\boldsymbol{x}} P(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{a}) \times \pi(\boldsymbol{a}|\boldsymbol{x}) D^{(l-1)}(\boldsymbol{x}) \quad \forall \boldsymbol{x}'$
end

Algorithm 1: Multi-Pedestrian Fictitious Play

Input : Initial state $\boldsymbol{x}_{0,n} \forall n$, τ
Output: Forecasted cumulative state visitation distribution $\{\bar{D}_n\}$

$D_n^{(0)}(\boldsymbol{x}_{0,n}) = 1$ for all n
for $t = \tau : \tau : T$ **do**
  for $n = 1 : N$ **do**
    $\mu_{\neg n}^{(t)} \leftarrow$ UpdateEmpirical$(\{\mu_m^{(t-\tau)} | m \neq n\})$ (Eq.1)
    $f_{n,soc}^{(t)} \leftarrow$ EncodeToFeature$(\mu_{\neg n}^{(t)}, D_{\neg n}^{(t-\tau)})$ (Alg.2)
    $U_n^{(t)} \leftarrow$ UpdateUtility$(f_{n,soc}^{(t)})$ (Eq.2)
    $D_n^{(t:t+L_n)} \leftarrow$ TakeMacroAction$(U_n^{(t)}, D_n^{(t-\tau)})$ (Alg.3)
    $\bar{D}_n = \bar{D}_n + \sum_{l=t}^{t+\tau-1} D_n^{(l)}$
  end
end

This process is repeated for T time steps.

UpdateEmpirical. Under the assumptions of fictitious play, the empirical distribution over opponent macro-actions $\mu_{\neg n}(\boldsymbol{s}_{\neg n})$ is typically computed by counting how many times each macro-action was chosen by each player. In our case, we need to describe a distribution over trajectories and so we use a parameterized form of the empirical distribution (*i.e.*, a maximum entropy distribution). The empirical distribution over macro-actions of all other pedestrians is decomposed into a product of distributions for each pedestrian $\mu_{\neg n}(\boldsymbol{s}_{\neg n}) \propto \prod_{m \neq n} \mu_m(\boldsymbol{s}_m)$. Each distribution is parametrized by a maximum entropy probability (also called Boltzmann or Gibbs) distribution,

$$\mu_m(\boldsymbol{s}_m) \propto \exp \sum_{\boldsymbol{x} \in \boldsymbol{s}_m} \boldsymbol{\theta}^\top \boldsymbol{f}_m(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{f}_m(\boldsymbol{x})$ are the features of a state $\boldsymbol{x}$ along the trajectory $\boldsymbol{s}_m$ for the pedestrian m, which are weighted by the vector of parameters $\boldsymbol{\theta}$. We will explain in Section 3.2 how the parameters $\boldsymbol{\theta}$ of the empirical distribution are learned from a dataset of demonstrated pedestrian behavior.

EncodeToFeature. This function maps $\mu_{\neg n}^{(t)}$ to the feature vector $f_{n,soc}^{(t)}$. Intuitively, this function predicts how all other pedestrians will move in the next few time steps and converts that predicted distribution into a state feature. For each pedestrian m, we compute their state visitation distribution $D_m^{(t:t+L_m)}$, which describes the likelihood of pedestrian m being in a certain location at a certain time step. The state visitation distributions of all other pedestrians $\bar{D}_{\neg n}$ are then summed together to generate the state feature $f_{n,soc}^{(t)}$.

UpdateUtility. In order to predict how each pedestrian will move over a sequence of time steps, and to compute how those predictions will affect the predictions of other pedestrian, we need to use a time-varying utility function for each pedestrian n,

$$U_n^{(t)}[\boldsymbol{s}_n, \mu_{\neg n}^{(t)}(\boldsymbol{s}_{\neg n})] \propto \exp \sum_{\boldsymbol{x} \in \boldsymbol{s}_n} \boldsymbol{\theta}^\top \boldsymbol{f}_n^{(t)}(\boldsymbol{x}). \tag{2}$$

Notice that the utility function is also a maximum entropy distribution, where the empirical distribution of all other pedestrians $\mu_{\neg n}^{(t)}(\boldsymbol{s}_{\neg n})$ has been incorporated through the feature vector $\boldsymbol{f}_n^{(t)}(x)$ (details in Section 3.3). The utility function is updated every τ time steps, the frequency at which each pedestrian makes predictions about the movement of others. A distribution over states sampled from U_n in illustrated in Figure 3 (right).

It is important to make a connection between the utility function U and the empirical distribution μ at this juncture. In our formulation, U_m is exactly equivalent to $\mu_m(\boldsymbol{s}_m)$. In general, U need not be a probability distribution, as it simply describes the value (reward) of one macro-action over another. In contrast, the empirical distribution μ is a probability distribution by construction and it describes which macro-action the opponent is likely to take. More importantly, the utility function U helps us to understand the dependency of the predicted path of pedestrian n on the predicted path of all other pedestrians $\neg n$. When we use the utility function to forecast the path of a single agent, that prediction influences the predicted path of all other predictions. This interplay between forecasted paths is precisely what we set out to model.

TakeMacroAction. This function takes the current empirical distribution $\mu_n^{(t)}$ and the prior state visitation distribution $D_n^{(t-1)}$ to compute the future state visitation distribution $D_n^{(t:t+L_n)}$. Intuitively, this function forward simulates pedestrian motion L_n steps into the future. To compute the future state visitation distribution $D_n^{(t:t+L_n)}$, a policy

π is first derived from the empirical distribution (this process described in Section 3.2). Using that policy, we iteratively compute how the prior state distribution $D_n^{(t-1)}(\boldsymbol{x})$ will change in over the next L_n time steps.

As a more concrete example, Figure 3 illustrates the procedure where we employ fictitious play to model the interactions within three pedestrians. The three pedestrians, respectively colored in red, green, and yellow, sequentially make predictions (*i.e.*, fictitious play) of others' macro-actions based on $\mu_n(\boldsymbol{s}_{\neg n})$ and then take the macro-action that maximize one's utility function. The forecasted state visitation distribution $\bar{D}_n$ (detailed in Section 3.3) of each pedestrian is expressed in the corresponding color and grows incrementally over time.

3.2. A Decision-Theoretic Pedestrian Model

We now explain how to learn the maximum entropy distribution which is used for both the utility function U_n and independent empirical distributions μ_m. As we have alluded to earlier, the probability of generating a trajectory $\boldsymbol{s}$ is modeled to be drawn from a maximum entropy distribution, where the probability is proportional to the exponentiated sum of weighted features encountered over the trajectory,

$$P(\boldsymbol{s};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \sum_{\boldsymbol{x}\in\boldsymbol{s}} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}), \tag{3}$$

where Z is the normalization function (or partition function), $\boldsymbol{\theta}$ is a vector of parameters and $\boldsymbol{f}(\boldsymbol{x})$ is a vector of features at state $\boldsymbol{x}$.

In order to learn the parameters $\boldsymbol{\theta}$ of this model from a set of demonstrated pedestrian trajectories, we utilize maximum entropy inverse optimal control [48]. We first make the assumption that each pedestrian is a rational agent and plans a path according to an underlying Markov Decision Process (MDP). The MDP describing a pedestrian n is defined by an initial state distribution $P_n(\boldsymbol{x}_0)$, a transition model $P_n(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{a})$ and a reward function $R_n^{(t)}(\boldsymbol{x})$. Following [21], the reward function is further defined as a weighted combination of features, $R_n^{(t)}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{f}_n^{(t)}(\boldsymbol{x})$. Note however that our reward function $R_n^{(t)}(\boldsymbol{x})$ is time indexed, as the feature vector $\boldsymbol{f}_n^{(t)}(x)$ will be used to encode information about changes in predicted behaviors of other pedestrians.

To learn the parameters $\boldsymbol{\theta}$ using maximum entropy IOC [48], we implement a gradient descent procedure that first computes a policy $\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})$ based on the current estimate of $\boldsymbol{\theta}$. The we compute the gradient update using difference between the estimated cumulative feature count and empirical cumulative feature count over demonstrated trajectories given that policy. When the features accumulated over trajectories generated by the MDP model converge to values similar to the empirical feature counts of the training data

Figure 4. The green box demonstrates how a pedestrian forms beliefs about others $\mu_{\neg n}^{(t)}$ and encode such information into social compliance feature $f_{n,soc}^{(t)}(\boldsymbol{x})$. Red indicates high reward and blue indicates low reward.

(*i.e.*, likelihood under the maximum entropy distribution is maximized), the algorithm has obtained an optimal set of parameters $\hat{\boldsymbol{\theta}}$, which will be used to define the empirical distribution μ.

An optimal policy for the maximum entropy distribution $P(\boldsymbol{s};\boldsymbol{\theta})$ can be computed as $\pi(\boldsymbol{a}|\boldsymbol{s}) = \exp\{Q(\boldsymbol{x},\boldsymbol{a}) - V(\boldsymbol{x})\}$, where the state-action soft value function $Q(\boldsymbol{x},\boldsymbol{a})$ and state soft-value function $V(\boldsymbol{x})$ can be computed by iterating the soft-maximum Bellman update equations: $Q(\boldsymbol{x},\boldsymbol{a}) = \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}) + E_{P(x'|x,a)}[V(\boldsymbol{x}')]$ and $V(\boldsymbol{x}) = \text{softmax}_a Q(\boldsymbol{x},\boldsymbol{a})$. We call this procedure ComputePolicy in Algorithm 2. Recall that in our scenario, the policy is time-varying since the features of the states change over time. Therefore, the policy for each pedestrian must be recomputed each time features are updated, *i.e.*, every forecasting period τ.

3.3. Features for Forecasting

In this section, we first show how the policy can be used to design the time dependent social compliance feature $f_{n,soc}^{(t)}$, which captures the interdependent reasoning process among people by encoding the empirical distribution over trajectories of all other agents $\mu_n^{(t)}(\boldsymbol{s}_{\neg n})$ into it. Then we build upon prior work [21] and introduce the semantic scene features which encode the intuition that rational agents will take into account the physical layout of the scene as they plan their future trajectories. Finally, we estimate the initial body orientation of each agent to encourage predictions that are aligned with body directions.

Social Compliance Feature: Given a policy $\pi(\boldsymbol{a}|\boldsymbol{x})$, we can generate a state visitation distribution D_n of pedestrian n for trajectories of length L_n by recursively computing:

$$D_n^{(l)}(\boldsymbol{x}') = \sum_{\boldsymbol{a},\boldsymbol{x}} P(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{a})\pi(\boldsymbol{a}|\boldsymbol{x})D_n^{(l-1)}(\boldsymbol{x}), \tag{4}$$

where $D_n^{(0)}(\boldsymbol{x})$ needs to be initialized to a distribution over start locations. Since $D_n^{(l)}(\boldsymbol{x})$ is defined over the entire state space, it is the same size as the state space. We can sum visitation counts over time, $\bar{D}_n(\boldsymbol{x}) = \sum_l D_n^{(l)}(\boldsymbol{x})$ to generate a cumulative distribution over states. The cumulative state visitation distribution $\bar{D}_n(\boldsymbol{x})$ represents the states that are likely to be occupied by pedestrian n when sampling from the empirical distribution $\mu_n(\boldsymbol{s}_n)$. By aggregating the cumulative visitation distribution for all pedestrian except n,

we can obtain a predicted occupancy map of all pedestrians in the environment, $\bar{D}_{\neg n}(\boldsymbol{x}) = \sum_{m \neq n} \bar{D}_m(\boldsymbol{x})$, which we will use to form our social compliance feature $f^{(l)}_{n,soc}$. Formally, this quantity encodes the empirical distribution $\mu_{\neg n}(\boldsymbol{s}_{\neg n})$ passed to the utility function (Equation 2). The process is summarized in Algorithm 2.

More intuitively, this quantity describes a social force field. Based on Helbing and Molonar's model of social forces [13] we define several social distance features that places a force field of varying size around the predicted trajectories of all other pedestrians in the environment. In particular, we defined three different sized force-fields, that roughly corresponds to Hall's proxemics zones [12], to encode a range of physical distances that people may maintain when walking in crowded scenes. Note that the social compliance features $f^{(t)}_{soc}(\boldsymbol{x})$ are indexed by time as the predicted paths of other pedestrians changes over time. It is also interesting to note that our feature naturally supports group behavior analysis, even though we do not explicitly model it as in [45]. When summing over the cumulative distribution of all other pedestrians, the states nearby groups will have larger visitation counts (*i.e.* more likely to be occupied), resulting in a more collision prone area. Figure 4 shows a situation where the area in front of the couple is a high potential collision area and thus has a lower reward.

Neighborhood Occupancy: This feature is a measurement of the amount of obstacles in a local neighborhood around a certain state. We calculate the number of pixels labeled as obstacles in a 5×5 grid and normalize it to provide a soft estimate of whether a state is a obstacle or not. The feature encodes how close pedestrians will walk near static objects in the scene. The neighborhood occupancy feature is denoted as $f_{occ}(\boldsymbol{x})$ which is not time varying as we assume the geometry of the scene to be static.

Distance-to-Goal: The feature $f_{dog}(\boldsymbol{x})$ captures a pedestrian's desire to approach his goal quickly by computing the Euclidean distance between a state $\boldsymbol{x}$ and the goal $\boldsymbol{x}_g$.

Body Orientation: Since a pedestrian's body orientation is a strong cue of the direction in which she will walk [10], we train a CNN (described in details in Section 3.4) to predict the initial walking direction of a pedestrian. We use the value of cosine distance minus one over the 8 connected neighbors centered at current pedestrian location. The value is greatest (0) in the direction of the predicted velocity direction and is the lowest (-2) in the opposite direction. The body orientation feature is denoted as $f_{bod}(\boldsymbol{x})$.

3.4. Walking Characteristics from Appearance

We further enhance the predictive power of our multi-pedestrian framework by allowing the model to maintain individualized walking models for each pedestrian based on appearance. In this section, we focus on visual information which conveys salient cues about how each individual in the scene may walk. For example, when we walk in crowds, the initial body orientation of a person may inform us of which direction that individual might walk. We may also perform high-level visual inference, predicting that an elderly couple might walk slow or a young business man might walk with a brisk pace. We propose using visual classifiers to identify various attributes of a pedestrian, and then map those attributes to walking direction and speed.

Figure 5. An overview of the datasets we used in the experiments.

To extract attributes from a pedestrian's visual appearance, we make use of a deep learning model. In particular, we employ a network structure similar to [23], but modify the top layer to generate three classification outputs: (1) age (old or young), (2) gender (male or female) and (3) body orientation (8 discretized direction). We train all three top layer classifiers jointly, as previous work has shown that multi-task learning helps to constrain the parameter learning [39, 47]. The predicted body orientation is used to generate the body orientation feature mentioned in Section 3.3, while the output of the age and gender classifiers are used to build individualized pedestrian models.

To be concrete, we use the soft probabilistic output of the age and gender classifiers to estimate an individualized velocity parameter. For each pedestrian n, we compute the individual's velocity v_n as the weighted average over gender and age velocity averages, *i.e.* $v_n = \sum_a w_a v_a^{stats}$, where $a \in \{male, female, old, elder\}$ denotes the attributes, w_a denotes the softmax output from deep net, and v_a^{stats} represents the average speed of pedestrians with attribute a. The individualized speed v_n is then incorporated into our model by multiplying the forecasting window size W, *i.e.* $L_n = W \times v_n$. Recall that L_n is the length of macro-actions s_n and v_n denotes speed, W can thus be interpreted as *how many time steps into the future one will predict about others*. In general, given a fixed W, the faster a pedestrian walks, the larger his occupancy map $\bar{D}_n = \sum_{l=t}^{t+W*v_n} D_n^{(l)}$ may be. We note that when speed information is not available, we employ a constant speed C for every pedestrian, *i.e.*, $L_n = W \times C \ \forall n$. We also tried regressing velocity directly from appearance, but in practice deep nets fail to learn discriminative features for direct regression.

4. Experiments

We analyze our model from various aspects. Following [21], we first assume the destinations are known to evaluate our forecasting performance in isolation. To validate the effectiveness of our model in the real world, we later perform unconstrained experiments with unknown goals.

NLL	nMDP[21]	MDPCV	mTA[30]	FP	FP + Speed
Zara [25]	46.5396	46.9549	43.3834	**42.1426**	-
Town Centre [5]	14.4797	14.4011	14.2471	12.5804	**10.892**
LIDAR Trajectory	92.5579	93.1747	91.9748	**87.4680**	-
Zara (no-dest) [25]	98.7343	97.6634	92.8271	**88.5693**	-
Town Centre (no-dest) [5]	33.8454	33.3213	31.5433	27.5732	**27.2136**
LIDAR Trajectory (no-dest)	173.643	175.384	169.782	**161.338**	-

SCR	nMDP[21]	MDPCV	mTA[30]	FP	FP + Speed
Zara [25]	0.144	0.114	0.065	**0.013**	-
Town Centre [5]	0.215	0.213	0.120	0.052	**0.049**
LIDAR Trajectory	0.133	0.105	0.056	**0.009**	-
Zara (no-dest) [25]	0.186	0.175	0.095	**0.021**	-
Town Centre (no-dest) [5]	0.323	0.281	0.170	0.093	**0.066**
LIDAR Trajectory (no-dest)	0.197	0.173	0.082	**0.022**	-

Table 1. Comparative analysis between different approaches. Smaller values are better.

We evaluate our model on three different pedestrian interaction datasets: the Zara Dataset [25], the Town Centre Dataset [5], and the LIDAR Trajectory Dataset. The first two datasets represent real world crowded settings with non-linear trajectories. To show that our model can also work with other modes of trajectory data, we further collected a LIDAR-based Trajectory Dataset, consisting of 20 interactive trajectories. Subjects are initialized at various start locations in a small room ($7m \times 7m$) with a few obstacles. They are then directed to walk towards a goal location without colliding with other pedestrians in the scene. We show samples of each dataset in Figure 5.

4.1. Metrics and Baselines

Negative Log Loss (NLL). The negative log loss computes the likelihood of drawing the demonstrated trajectory. It is defined as $NLL(\boldsymbol{s}) = -\sum_t \log \pi^{(t)}(\boldsymbol{a}^{(t)}|\boldsymbol{x}^{(t)})$, where trajectory $\boldsymbol{s}$ is a sequence of state-action pairs $(\boldsymbol{x}, \boldsymbol{a})$.

State Collision Rate (SCR). While the NLL is appropriate for evaluating single agent forecasting results, it does not explicitly penalize colliding predicted paths in the multi-agent forecasting case. To encode the notion of a future collision, we define the State Collision Rate $SCR = \sum_t \prod_n D_n^{(t)}(\boldsymbol{x})$, where n denotes a pedestrian ID, and $D_n^{(t)}(\boldsymbol{x})$ represents the expected state visit count at state $\boldsymbol{x}$ at time t, *i.e.*, the probability of being at certain state at a certain time. By taking into account the distribution of multiple pedestrians and taking their union, the resulting state visit count of all agents represents regions of collision.

We compare with the following three baselines:

N-Independent MDP (nMDP). This baseline model is the approach of [48] applied to images for forecast the trajectory of a single pedestrian [21]. Extending their approach for our multi-agent scenario, we use N instantiations of their MDP model and run them in parallel.

MDP + Constant Velocity (MDPCV). The second baseline model is a modification of the Independent MDP model but with a collision region features added to the reward function. By assuming constant velocity, we can compute regions of collision (*i.e.*, the intersection regions of linear motion models) and encode them using the same way as the neighborhood occupancy feature.

mTA. Based on the work of Pellegrini *et al.* [30], the third approach is a modified Trajectory Avoidance (mTA) model. In [30] every agent chooses a velocity that minimizes its energy function at every time step. Formulated as an MDP, this corresponds to a reward function using only a constant feature. As for modeling the social force features (*e.g.*, comfortable distance among agents), we use the social compliance features described in Section 3.3. We emphasize here that this baseline model has less information that the original model described in [30] where every agent knows the positions and velocities of others. This information is not available in our problem setup (*i.e.*, single image input).

	Young	Old	Male	Female
Average Speed (grids/frame)	1.98	1.25	1.78	1.53

	Age	Gender	Body Direction
Accuracy	82.31%	78.44%	65.60%

Table 2. Top: average speed of people with different visual attributes. Bottom: accuracy of visual pedestrian classification.

4.2. Multi-Pedestrian Forecasting Performance

To properly evaluate our proposed approach, we apply our method only on trajectory sequences that demonstrate strategic reasoning, where multiple pedestrians are actively avoiding each other as they walk. We obtain 16 multi-pedestrian trajectory sequences from each dataset [25] and [5][1]. We emphasize here that trajectories of single pedestrians simply walking in a straight line are not used, as it is possible to artificially increase performance by adding more of these 'easy' examples. We compute the metrics with 5-fold cross validation. The forecasting window size is set to $W = 3$ and the forecasting period is $\tau = 1$. The two parameters are found via grid search, and a detailed analysis can be found in the supplementary material. Results are summarized in Table 1. We observe that our fictitious play based approach outperforms all three approaches with respect to NLL and SCR. This shows that our iterative predicting and planning process better predicts human interactions and also generates the most collision-free trajectories.

We further incorporate speed information into our model using the method mentioned in Section 3.4. We evaluate the effectiveness of speed information on the Town Centre dataset as the resolution of the Zara dataset is too low to extract visual features and the LIDAR-based Trajectory dataset does not provide any visual features. We collected ≈16K pedestrian patches from the Town Centre Dataset [5], with three labels for each patch, *i.e.*, age, gender, and body orientation. The images are split into the corresponding 5-fold by pedestrians. We train a deep classifier using the network structure in Section 3.4. The performance is shown in Table 2(bottom). We also computed speed statistics from

[1]For more details in how we select the trajectories and the results on the original dataset, please refer to the supplementary material.

Figure 6. Multi-agent forecasting examples of pre-emptive collision avoidance. Each pedestrian is marked with a colored bounding box, with corresponding forecasting distribution in the same color. Note that we consider all pedestrians for quantitative experiments but only visualize forecast distributions for a limited number of pedestrian to improve visualization.

the videos (see Table 2(top)). Using the speed statistics and the output of the deep network model, we can compute an individualized model for each pedestrian. As expected, Table 1 shows that our model performs better after considering a pedestrian's visual appearance and individualize the predictive model. Selected qualitative results of predicted trajectories are shown in Figure 6.

Relax Destination Constraints. To show that our model also work in the real work settings where the final destination of a pedestrian is not known and needs to be inferred, we follow [21] to densely generate potential goals on the map and perform the same forecasting experiment. Results, denoted as *no-dest* in Table 1, show that our FP based approach, which models the interplay and visual evidence, still consistently outperforms others even without knowing the destinations ahead of time. The absolute performance of all models degrade due to uncertainty about the goal.

4.3. Features Analysis

We further evaluate the effects of the features used in our proposed model. The average NLL and SCR for the Town Centre dataset using different features are shown in Table 3 (results on other datasets can be found in the supplementary material). We set forecasting window size $W = 3$ and forecasting period $\tau = 1$ as before. The performance of other approaches are also shown for reference. Note that nMDP and MDPCV still consider only scene features and body orientation features even with the social compliance feature checked, since they cannot handle the dynamics in the environment. Our model performs identical to nMDP when considering only semantic scene features and body orientation features. This result is expected as there are no social compliance features to change the cost topology over time. If there is only one agent (which implies there is no social compliant feature), our proposed model reduces to nMDP. We emphasize that with the inclusion of the social compliance feature, our proposed models better explains the interactions between multiple pedestrians. The FP+Speed model attains a NLL of 10.892 compared to the next best performing model mTA at 14.247, resulting in a 23.5% improvement in the NLL.

	f_{occ}	f_{dog}	f_{bod}	f_{soc}	nMDP[21]	MDPCV	mTA[30]	FP	FP + Speed
NLL	✓	✓			14.48	14.40	14.48	14.48	14.48
	✓	✓	✓		14.45	14.33	14.45	14.45	14.45
	✓	✓		✓	14.48	14.40	14.29	12.65	11.01
	✓	✓	✓	✓	14.45	14.33	14.25	12.58	**10.89**
	f_{occ}	f_{dog}	f_{bod}	f_{soc}	nMDP[21]	MDPCV	mTA[30]	FP	FP + Speed
SCR	✓	✓			0.215	0.183	0.215	0.215	0.215
	✓	✓	✓		0.211	0.175	0.211	0.211	0.211
	✓	✓		✓	0.215	0.183	0.129	0.046	0.044
	✓	✓	✓	✓	0.211	0.175	0.120	0.043	**0.039**

Table 3. Contribution of each feature to our model.

5. Conclusion

We present a novel framework to forecast multi-pedestrian trajectories from a *single image* by directly modeling the interplay between multiple people using concepts from game theory and optimal control. We also develop various predictive models to show how different modes of information help to reason about the future actions of multi-pedestrian scenarios. By building individualized pedestrian models for each person based on his visual appearance, we generate more accurate prediction of multi-pedestrian interactions. We have compared our Fictitious Play based approach with other state-of-the-art algorithms. Our evaluation on multiple pedestrian interaction datasets has shown that our proposed approach is able to attain more accurate long-term predictions of pedestrian activity.

Acknowledgement This work was supported by JST CREST Grant Number JPMJCR14E1, Japan.

References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, 2016. 2

[2] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-aware large-scale crowd forecasting. In *CVPR*, 2014. 1, 2

[3] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*, 2007. 1, 2

[4] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *ECCV*, 2008. 1, 2

[5] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, 2011. 7

[6] S. Z. Bokhari and K. M. Kitani. Long-term activity forecasting using first-person vision. In *ACCV*, 2016. 2

[7] S. J. Brams. *Game theory and politics*. 2011. 3

[8] G. W. Brown. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 1951. 2, 3

[9] B. Cancela, A. Iglesias, M. Ortega, and M. G. Penedo. Unsupervised trajectory modelling using temporal information via minimal paths. In *CVPR*, 2014. 2

[10] I. Chamveha, Y. Sugano, D. Sugimura, T. Siriteerakul, T. Okabe, Y. Sato, and A. Sugimoto. Appearance-based head pose estimation with scene-specific adaptation. In *ICCV*, 2011. 6

[11] H. Gong, J. Sim, M. Likhachev, and J. Shi. Multi-hypothesis motion planning for visual object tracking. In *ICCV*, 2011. 2

[12] E. T. Hall. The hidden dimension. *Garden City*, 1966. 6

[13] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 1995. 1, 2, 6

[14] D.-A. Huang, A. M. Farahmand, K. M. Kitani, and J. A. Bagnell. Approximate maxent inverse optimal control and its application for mental simulation of human interactions. In *AAAI*, 2015. 2

[15] D.-A. Huang and K. M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. In *ECCV*. 2014. 2

[16] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In *ICCV*, 2015. 2

[17] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. *arXiv*, 2015. 2

[18] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto. Intent-aware long-term prediction of pedestrian motion. *ICRA*, 2016. 2

[19] K. Kim, M. Grundmann, A. Shamir, I. Matthews, J. Hodgins, and I. Essa. Motion fields to predict play evolution in dynamic sport scenes. In *CVPR*, 2010. 2

[20] S. Kim, S. J. Guy, W. Liu, D. Wilkie, R. W. Lau, M. C. Lin, and D. Manocha. Brvo: Predicting pedestrian trajectories using velocity-space reasoning. *The International Journal of Robotics Research*, 2014. 2

[21] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *ECCV*, 2012. 2, 5, 6, 7, 8

[22] H. Kretzschmar, M. Kuderer, and W. Burgard. Learning to predict trajectories of cooperatively navigating agents. In *ICRA*, 2014. 2

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 6

[24] N. Lee and K. M. Kitani. Predicting wide receiver trajectories in american football. In *WACV*, 2016. 2

[25] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In *Computer Graphics Forum*, 2007. 7

[26] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009. 2

[27] R. B. Myerson. *Game theory*. 2013. 3

[28] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic game theory*. 2007. 3

[29] H. S. Park, E. Jain, and Y. Sheikh. 3d social saliency from head-mounted cameras. In *NIPS*, 2012. 2

[30] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009. 1, 2, 7, 8

[31] M. Rabin. Incorporating fairness into game theory and economics. *The American economic review*, 1993. 3

[32] N. Rhinehart and K. M. Kitani. Online semantic activity forecasting with darko. *arXiv*, 2016. 2

[33] M. Ryoo, T. J. Fuchs, L. Xia, J. K. Aggarwal, and L. Matthies. Robot-centric activity prediction from first-person videos: What will they do to me'. In *HRI*, 2015. 2

[34] P. Scovanner and M. F. Tappen. Learning pedestrian dynamics from the real world. In *ICCV*, 2009. 2

[35] O. Sener and A. Saxena. rcrf: Recursive belief estimation over crfs in rgb-d activity videos. In *RSS*, 2015. 2

[36] H. Soo Park, J.-J. Hwang, Y. Niu, and J. Shi. Egocentric future localization. In *CVPR*, 2016. 2

[37] B. Soran, A. Farhadi, and L. Shapiro. Generating notifications for missing actions: Don't forget to turn the lights off! In *ICCV*, 2015. 2

[38] B. Tastan and G. Sukthankar. Leveraging human behavior models to predict paths in indoor environments. *Pervasive and Mobile Computing*, 2011. 2

[39] Y. Tian, P. Luo, X. Wang, and X. Tang. Pedestrian detection aided by deep learning semantic tasks. *arXiv*, 2014. 6

[40] C. Vondrick, H. Pirsiavash, and A. Torralba. Anticipating the future by watching unlabeled video. *arXiv*, 2015. 2

[41] J. Walker, A. Gupta, and M. Hebert. Patch to the future: Unsupervised visual prediction. In *CVPR*, 2014. 2

[42] J. Walker, A. Gupta, and M. Hebert. Dense optical flow prediction from a static image. *arXiv*, 2015. 2

[43] D. Xie, T. Shu, S. Todorovic, and S.-C. Zhu. Modeling and inferring human intents and latent functional objects for trajectory prediction. *arXiv*, 2016. 2

[44] D. Xie, S. Todorovic, and S.-C. Zhu. Inferring" dark matter" and" dark energy" from videos. In *ICCV*, 2013. 2

[45] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg. Who are you with and where are you going? In *CVPR*, 2011. 6

[46] S. Yi, H. Li, and X. Wang. Understanding pedestrian behaviors from stationary crowd groups. In *CVPR*, 2015. 2

[47] X.-T. Yuan, X. Liu, and S. Yan. Visual classification with multitask joint sparse representation. *TIP*, 2012. 6

[48] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, 2008. 5, 7

[49] A. Zunino, J. Cavazza, A. Koul, A. Cavallo, C. Becchio, and V. Murino. Intention from motion. *arXiv*, 2016. 2