A Game-Theoretic Approach to Multi-Pedestrian Activity Forecasting

Wei-Chiu Ma

CMU-RI-TR-16-16
May 2016

The Robotics Institute
School of Computer Science
Carnegie Mellon University

Thesis Committee:
Kris M. Kitani, Chair
Deva Ramanan
Wen-Sheng Chu

Submitted in partial fulfillment of the requirements for the degree of
Master of Science in Robotics

Keywords: Activity Forecasting, Fictitious Play, Multi-Pedestrian Modeling

For my grandpa

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. 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 predicted behavior of one pedestrian can affect the actions taken by others and (2) dynamics are variable depending on an individuals physical characteristics (*e.g.*, an older person may walk slowly while the younger person may walk faster). To address these challenges, we (1) utilize concepts from game theory to model the interdependent decision making process of multiple pedestrians and (2) 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 explains human interactions 25% better when compared to state-of-the-art methods.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Kris M. Kitani, who led me into the field of computer vision and taught me how to conduct scientific research. His patience, guidance and knowledge has been invaluable in my work. Without his continuous support, this thesis would not have been possible. To me, Kris is not only an advisor, but also an inspiring mentor.

I'm also extremely grateful to all my fellows, especially De-An Huang, Chao Liu, Minghuang Ma, and Wen-Sheng Chu. Many part of the thesis grew out from countless conversations with them.

Last but not least, I would like to thank my beloved parents, sister, and grandparents for their unconditional support and endless love.

Contents

1 Introduction **1**

2 Related Work **5**

3 Forecasting Multi-Pedestrian Trajectories **7**
3.1 Forecasting Interactions as Fictitious Play 7
3.2 A Decision-Theoretic Pedestrian Model 10
3.3 Features of the Reward Function 11
3.4 Walking Characteristics from Appearance 13

4 Experiments **15**
4.1 Metrics 16
4.2 Multi-Pedestrian Forecasting Performance 17
4.3 Forecasting Window Size and Forecasting Period 19
4.4 Features for Forecasting 20
4.5 Destination Forecasting 21

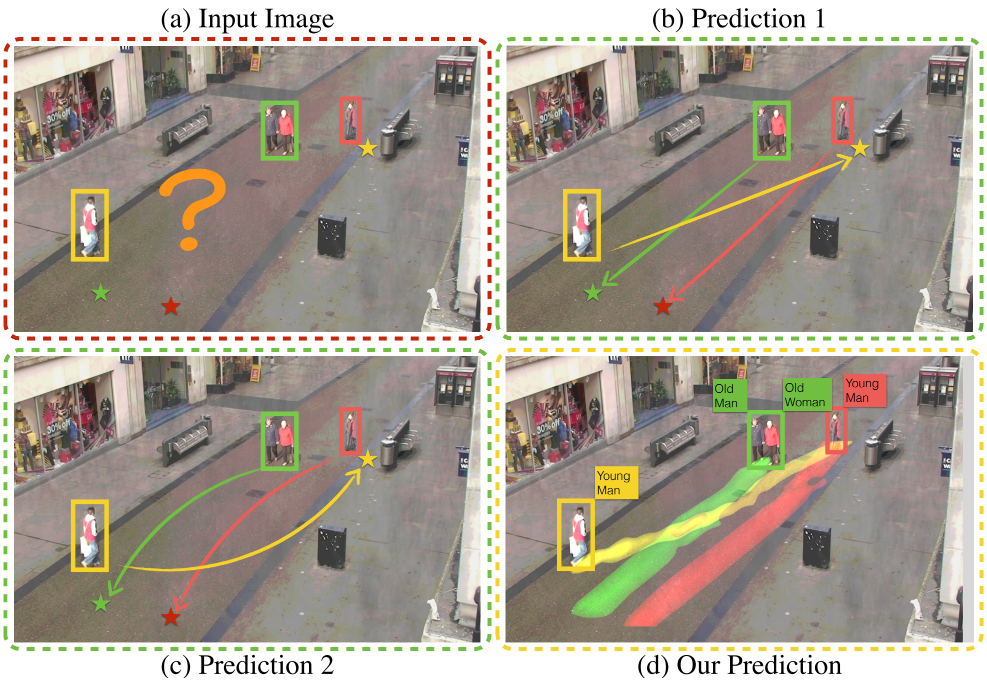
5 Conclusion **23**

Bibliography **25**

Chapter 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 as depicted in Figure 1.1, 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 1.1(b).This strategy, however, might lead to collisions between pedestrians (*e.g.*, the young man (yellow) and the elder couple (green) may collide). A more thoughtful model might consider the possibility


(a) Input Image

(b) Prediction 1

(c) Prediction 2

(d) Our Prediction

Figure 1.1: Given the top left image where pedestrians are surrounded by colored bounding boxes with their corresponding destinations marked with colored star signs, can you forecast their future behavior? What can you tell from this single image?

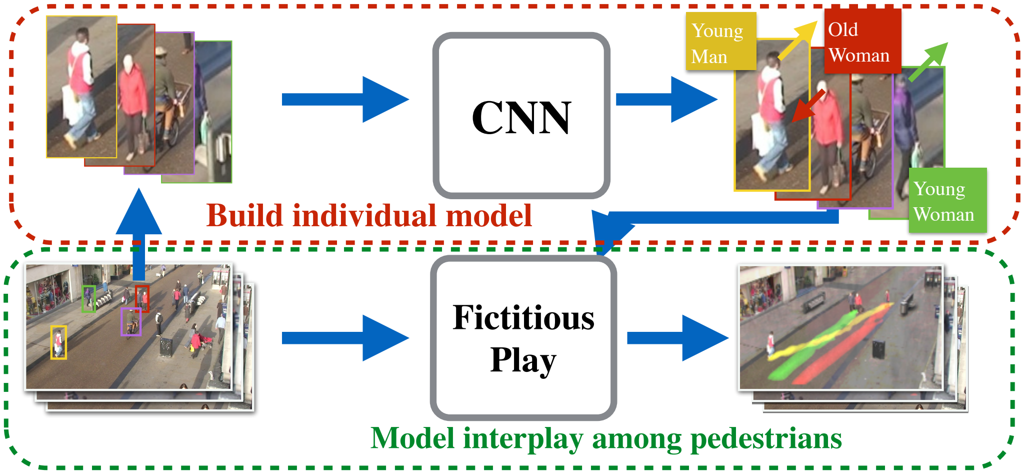


Figure 1.2: **Overview of our approach.** Given a single image and pedestrian detections, we estimate physical properties for each pedestrian using deep neural network. Based on estimated properties, we forecast multi-pedestrian interactions using Fictitious Play.

that one or more pedestrians will alter their trajectory based on their *prediction of other pedestrians*. A possible trajectory set is shown in Figure 1.1(c), where each pedestrian demonstrates preemptive collision avoidance. 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 to have a brisk walk. In this case, by considering velocity adjustments, we can reason that all pedestrians will roughly walk straight to their destination. This illustration serves to highlight the complex reasoning that is involved in predicting the walking trajectories of several people. Our goal is to mimic – computationally – this ability to reason about the dynamics of interactive social processes.

Developing computational models for forecasting human activities is an important feature that autonomous system must possess in order to seamlessly interact with their human counterparts. For example, activity forecasting is an important feature needed for domestic robotic systems which need to be able to predict future activities of people to assist them or simply to stay out of their way (*e.g.*, collision avoidance). Autonomous cars must be able to predict the possible trajectories of pedestrians to plan safer routes. Activity forecasting can also be used to detect anomalous activity: by comparing observed human actions against predicted normative behavior, an autonomous system can understand when its human counterpart is doing something out of the ordinary. The realization of such predictive capabilities requires new models of human activity which can reason about complex futures.

In this work, 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 [6] 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 empirical distribution over the strategies of other players. As we will show, the multi-player game model has strong parallels to multi-pedestrian navigation, as each pedestrian pre-emptively plans his path according to his beliefs about how other pedestrians will move in the future.

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 pedestrians 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 1.2 shows the overview of our approach.

The contributions of this paper are as follows: (1) 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; and (2) we build individualized predictive pedestrian models to generate more accurate forecasts of multi-pedestrian interactions.

Chapter 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 [7, 12, 13, 14, 15, 19, 29, 30, 31, 32, 33, 35, 37]. In the context of pedestrian dynamics, Helbing and Molonar [11] 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 [27] and anomaly detection [23]. More recent work has focused on discovering the underlying potential field by observing human behavior such as patterns of motions [1], mutual 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 [2, 3] and the interaction between stationary crowds and pedestrians can be used to predict pedestrians' future motions [39]. 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 [16, 26]. Patterns of avoidance can also be used to learn the hidden rewards or costs of physical spaces [18, 36, 38].

To make reliable predictions about the long-term future, many techniques often assume a static environment [18, 36]. 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 [9, 17, 27] (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 dynamic environments remains relatively unexplored in visual human activity analysis except [15, 21]. In these work, the complex and intertwined interactions between agents is either ignored [15] or restricted to the perspective of a single agent (wide receiver in [21]). In contrast, 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.

Chapter 3

Forecasting Multi-Pedestrian Trajectories

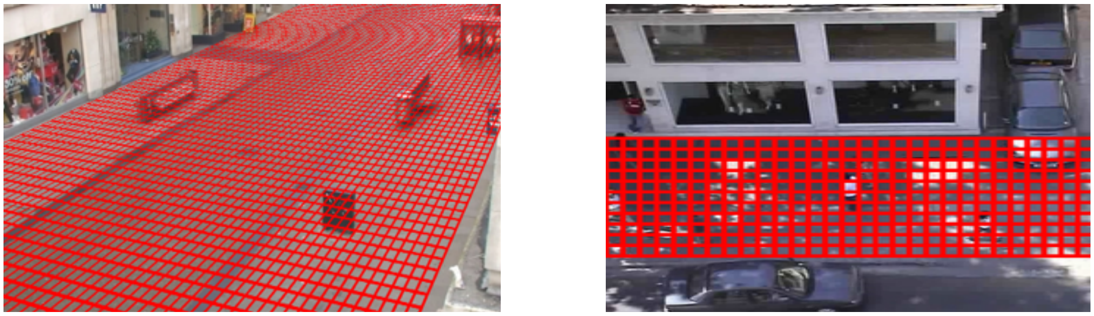
Given a single image and initial pedestrians detections, we aim to develop a predictive model that can forecast plausible trajectories for all pedestrians. To do this we must model the complex predictive interplay between multiple pedestrians, while also considering individual differences that impact 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 in deep neural network to build 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 and coordination system. We employ the same coordinate system as in [18]. We define the state space (the ground plane) as a 2D lattice, where each position is denoted by $\boldsymbol{x} = [x, y] \in X$. 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 states, $\boldsymbol{s} = \{x_1, \ldots, x_K\}$. We note that *all* computations are performed under this coordinates. For example, each state $\boldsymbol{x}$ has an associated vector of features $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}) \ldots f_J(\boldsymbol{x})]$, where $f_j(\boldsymbol{x})$ could represent properties of that state such as the output of a visual classifier, the distance to an object or predicted presence of another pedestrian. When computing speed statistics from the videos, we first project the locations of the pedestrians to our state space (3D floor plane), then compute their respective speed $\boldsymbol{a}$. The grid world representation is shown in Figure 3.1.

3.1 Forecasting Interactions as Fictitious Play

Game theory [24] 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 [28], politics [5] and computer science [25]. 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) [6], where we model each pedestrian to take a path based on his 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 of multiple pedestrians. Formally, each



(a) (b)

Figure 3.1: **Grid world representation.** The ground plane is discretized into cells which represent states. As our state space is the 3D floor plane and as such, 2D image features, trajectories are all projected to the floor plane (assuming camera parameters are known) for computations. (a) and (b) visualize how we discretize the Town Centre Dataset [4] and the Zara dataset [22] respectively.

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 $U_n : \boldsymbol{s}_n \rightarrow \mathbb{R}$ based on forecasted distributions over macro-actions of all other pedestrians $\mu_{\neg n}(\boldsymbol{s}_{\neg n})$. 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 3.2 (right).

Typically, in a repeated game, the empirical distribution over opponent macro-actions $\mu_{\neg n}(\boldsymbol{s}_{\neg n})$ is computed by counting how many times each macro-action was chosen by other players in the past. However in our model, we parametrize the empirical distribution in the following way. The empirical distribution over macro-actions of all other pedestrians can be decomposed into a product of distributions for each pedestrian: $\mu_{\neg n}(\boldsymbol{s}_{\neg n}) \propto \prod_{m \neq n} \mu_m(\boldsymbol{s}_m)$. This is called UPDATEEMPIRICAL in Algorithm 1. Each distribution is parametrized by a maximum entropy probability (also called Boltzmann or Gibbs) distribution,

$$\mu_m(\boldsymbol{s}_m) \propto \exp \sum_{\boldsymbol{x} \in \boldsymbol{s}_m} \boldsymbol{\theta}^\top \boldsymbol{f}_m(\boldsymbol{x}), \tag{3.1}$$

where $\boldsymbol{f}_m(x)$ 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 $\boldsymbol{\theta}$ can be learned from a dataset of demonstrated pedestrian behavior.

Now 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,

$$U_n^{(t)}[\boldsymbol{s}_n, \mu_{\neg n}^{(t)}(\boldsymbol{s}_{\neg n})] \propto \exp \sum_{\boldsymbol{x} \in \boldsymbol{s}_n} \boldsymbol{\theta}^\top \boldsymbol{f}_n^{(t)}(\boldsymbol{x}). \tag{3.2}$$

This is called UPDATEUTILITY in Algorithm 1. 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

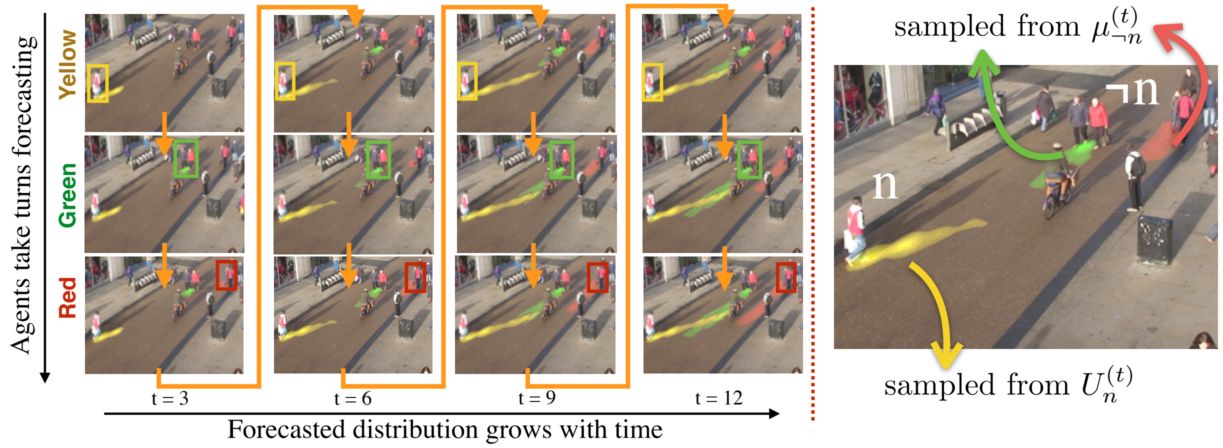

Agents take turns forecasting
Yellow
Green
Red
t = 3
t = 6
t = 9
t = 12
Forecasted distribution grows with time
sampled from $\mu_{\neg n}^{(t)}$
¬n
n
sampled from $U_n^{(t)}$

Figure 3.2: (Left) Visualization of Fictitious Play with three pedestrians. The orange arrows indicate the order. (Right) Distributions over states sampled from $U_n^{(t)}$ and $\mu_{\neg n}^{(t)}$.

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 make predictions about the movement of others. A distribution over states sampled from U_n in shown in Figure 3.2 (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 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.

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 $\mu_{\neg n}^{(t)}$ and encoding them into feature $f_{n,soc}^{(t)}$. To map $\mu_{\neg n}^{(t)}$ to the feature vector $f_{n,soc}^{(t)}$, we use the distribution $D_{\neg n}^{(t:t+L_n)}$, which is a function of states instead of macro-actions. $D_{\neg n}^{(t:t+L_n)}$ can be interpreted as the state visitation distribution and $f_{n,soc}^{(t)}$ is a element of $\boldsymbol{f}_n^{(t)}$ (details in 3.3). Each pedestrian updates his/her own utility function $U_n^{(t)}$ based on $\boldsymbol{f}_n^{(t)}$. The state visitation distribution $D_n^{(t:t+L_n)}$ is then drawn from the probability distribution over the macro-actions of pedestrian n and later encoded into the feature vector of other pedestrians in the next forecasting period.

Figure 3.2 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 D (detailed in Section 3.3) of each pedestrian is expressed in the corresponding color and grows incrementally forward with time.

```
Algorithm 1: Multi-Pedestrian Fictitious Play
Input   : Initial state $\boldsymbol{x}_{0,n} \forall n$, $\tau$
Output: Forecasted State Visitation Distribution $\{\bar{D}_n\}$
$\bar{D}_n^{(0)}(\boldsymbol{x}_{0,n}) = 1$ for all $n$
for $t = 1 : \tau : T$ do
    for $n = 1 : N$ do
        $\mu_{\neg n}^{(t)} \leftarrow$ UpdateEmpirical($\{\mu_m^{(t)} | m \neq n\}$)            (Eq.3.1)
        $f_{n,soc}^{(t)} \leftarrow$ EncodeToFeature($\mu_{\neg n}^{(t)}, D_{\neg n}^{(t-\tau)}$)            (Alg.2)
        $U_n^{(t)} \leftarrow$ UpdateUtility($f_{n,soc}^{(t)}$)            (Eq.3.2)
        $D_n^{(t:t+L)} \leftarrow$ TakeMacroAction($U_n^{(t)}, D_n^{(t-\tau)}$)            (Alg.3)
    end
end
```

```
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: $D_n^{(t:t+L_n)}$

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

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,

$$P(\boldsymbol{s};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \sum_{\boldsymbol{x} \in \boldsymbol{s}} \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}), \tag{3.3}$$

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 [41]. 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 [18], 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 the feature vector $\boldsymbol{f}_n^{(t)}(x)$ will be used to encode information about changes in the predicted behavior of other pedestrians.

To learn the parameters $\boldsymbol{\theta}$ using maximum entropy IOC [41], we implement a gradient descent procedure that first computes a policy $\pi(\boldsymbol{a}|\boldsymbol{s};\boldsymbol{\theta})$ based on the current estimate of $\boldsymbol{\theta}$ and then 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 the empirical feature counts of the training data (*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.

The optimal policy for a 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 following 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 of the Reward Function

As we have mentioned in Section 3.1, the feature vector $\boldsymbol{f}_n^{(t)}$ consists of two parts: time dependent features and time invariant features. We first show how the policy can be used to design the time dependent social compliance feature $f_{n,soc}^{(t)}$, which encodes the empirical distribution over trajectories for all other agents $\mu_n^{(t)}(\boldsymbol{s}_{\neg n})$. Then we explain the time invariant features in details.
Social Compliance Feature: For each pedestrian, we first compute the policy $\pi_n(\boldsymbol{a}|\boldsymbol{x})$ using the COMPUTEPOLICY procedure mentioned above. Given a policy $\pi_n(\boldsymbol{a}|\boldsymbol{x})$, we can then generate a state visitation distribution D_n of pedestrian n for trajectories of length L_n by recursively computing:

$$D_n^{(l)}(\boldsymbol{x}') = \sum_{\boldsymbol{a},\boldsymbol{x}} P(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{a})\pi_n(\boldsymbol{a}|\boldsymbol{x})D_n^{(l-1)}(\boldsymbol{x}), \tag{3.4}$$

where $D_n^{(0)}(\boldsymbol{x})$ needs to be initialized to a probability 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 that may be occupied by pedestrian n. 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 $\mu_n(\neg\boldsymbol{s})$. By aggregating the cumulative visitation distribution for all pedestrian except n, we can obtain a predicted occupancy map of all pedestrians in the environment, $\bar{D}_{\neg n}(\boldsymbol{x}) = \sum_{m\neq n} \bar{D}_m(\boldsymbol{x})$. Formally, this quantity encodes the empirical distribution $\mu_n(\boldsymbol{s}_{\neg n})$ passed to the utility function (Equation 3.2). The process is summarized in Algorithm 2 and 3.

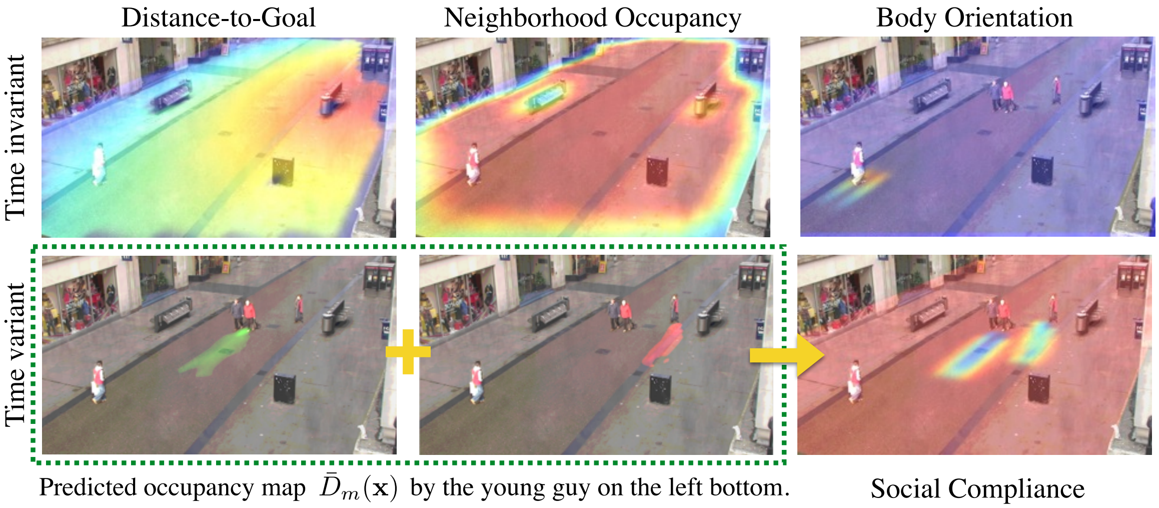


Figure 3.3: (Top) Features invariant of time. Neighborhood occupancy feature $f_{occ}(\boldsymbol{x})$ is used to encode how close a pedestrian will walk near static objects. Distance-to-goal feature $f_{dog}(\boldsymbol{x})$ describes pedestrians' desire to reach destination. Body orientation feature $f_{bod}(\boldsymbol{x})$ captures the fact that pedestrians tend to walk towards their body direction. (Bottom) The green box demonstrates how a pedestrian forms beliefs about others $\mu_{\neg n}^{(t)}$ and encode such information into social compliance feature $f_{soc}^{(t)}(\boldsymbol{x})$. Red indicates high reward and blue indicates low reward.

More intuitively, this quantity describes a social force field. Based on Helbing and Molonar's model of social forces [11] 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 [10], to encode a range of physical distances that people may maintain when walking in crowded scenes. The social compliance features $f_{soc}^{(t)}(\boldsymbol{x})$ is indexed by time as the predicted paths of other pedestrians changes over time.

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 estimation 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 he will walk, 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 $f_{bod}(\boldsymbol{x})$. Figure 3.3 visualize all features.

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 at brisk pace. We proposed to use visual classifiers to identify various attributes of a pedestrian, and then map those attributes to walking direction and speed.

To extract attributes from a pedestrians visual appearance, we make use of a deep learning model. In particular, we employ a network structure similar to [20], 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 [34, 40]. 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 his/her individual 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.* $L_n = W \times C$.

Chapter 4

Experiments

To validate our claim that (1) our fictitious play based approach better describes the interplay between pedestrians and (2) our deep learning based classifier can effectively extract features that are indicative of behavior pattern, we evaluate forecasting performance on three different pedestrian interaction datasets: (1) the Zara Dataset [22], (2) the Town Centre Dataset [4], and (3) the LIDAR Trajectory Dataset.

As our goal is to learn how people reason about the behavior of others and model the interactive dynamics among them, we focus on the video clips where pedestrians demonstrate *strategic reasoning*. We make an important assumption that people tend to take the shortest available path subjecting to the environmental constraints (*e.g.* pedestrians should walk on the sidewalk, people cannot cross the obstacles). If people explicitly adjust their path, we argue that this is affected by others. In particular, we employ the modified Hausdorff distance (MHD) to measure the physical distance between one's ground truth trajectory (provided by tracker) and the shortest possible trajectory (subjecting to environmental constraints). The MHD allows for local time warping by finding the best local point correspondence over a small temporal window and has been widely used to evaluate the similarity of two sequences [18]. The larger the MHD is, the more different the actual trajectory is to the shortest path, and the more obvious the interaction may be. The MHD is defined as:

$$\text{MHD}(\boldsymbol{s}_{gt}, \boldsymbol{s}_{short}) = \sum_{\boldsymbol{x}_{gt,n} \in \boldsymbol{s}_{gt}} \min_{w \in [n-W, n+W]} \|\boldsymbol{x}_{gt,n} - \boldsymbol{x}_{short,w}\|, \tag{4.1}$$

where $\boldsymbol{s}_{gt}$ and $\boldsymbol{s}_{short}$ respectively denote the ground truth trajectory and the shortest possible trajectory towards goal, $\boldsymbol{x}_{gt,n}$ and $\boldsymbol{x}_{short,w}$ refer to the nth and wth state in sequence $\boldsymbol{s}_{gt}$ and $\boldsymbol{s}_{short}$, W represents the small temporal window, and $\|\cdot\|$ denotes the euclidean distance function. If the MHD is larger than a certain threshold ϵ, we posit that there is an obvious interaction among pedestrians that leads to the change of trajectories. We thus include the trajectories of all pedestrians in the scene as one multi-pedestrian trajectory sequence. If the MHD is smaller than the threshold, we assume there is no obvious interactions and reject it. Following this criteria, we respectively obtain 16 multi-pedestrian trajectory sequences from the Zara Dataset [22] and the Town Centre Dataset [4]. Figure 4.1 illustrates our selection process.

To show that our model can also work with other modes of trajectory data, we introduce a LIDAR-based Trajectory Dataset. This dataset consists of 20 interactive trajectories satisfying

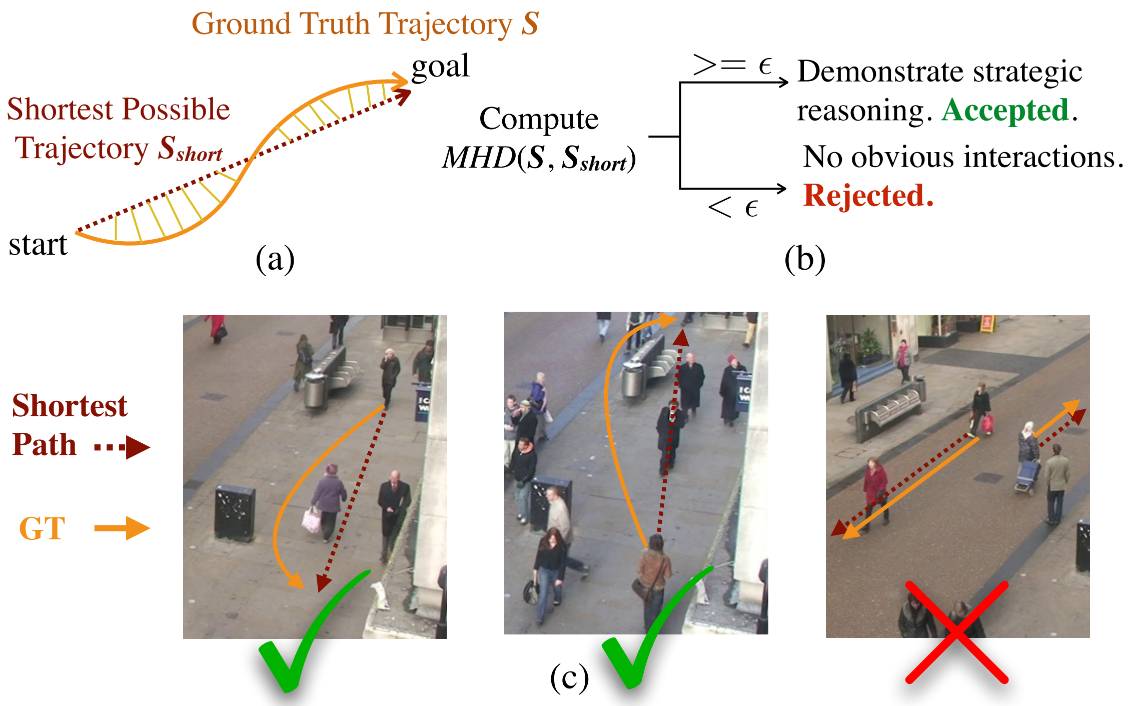

Ground Truth Trajectory S
goal
Shortest Possible Trajectory S_{short}
start
(a)
Compute $MHD(S, S_{short})$
$>= \epsilon$
Demonstrate strategic reasoning. Accepted.
No obvious interactions. Rejected.
$< \epsilon$
(b)
Shortest Path
GT
(c)

Figure 4.1: **Criteria of selecting training sequences for multi-pedestrian forecasting:** (a)(b) For every pedestrian in the scene, we compute the modified Hausdorff distance (MHD) between the ground truth trajectory and the shortest possible trajectory. If the MHD is larger than a certain threshold, we include the trajectories of all pedestrians in the scene as one multi-pedestrian trajectory sequence. If the MHD is smaller than the threshold, we reject it. (c) Examples of accepted and rejected seqences.

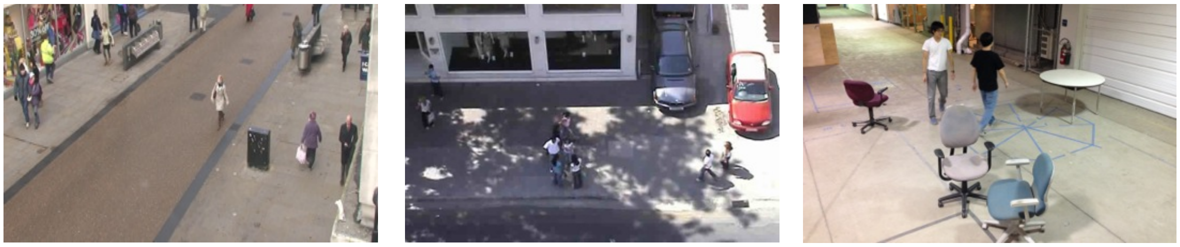
the criteria mentioned above. Subjects are initialized at various start locations in a small room ($7m \times 7m$) with a few obstacles. They are then directed to walk towards a goal location without colliding with other pedestrians in the scene. We show samples of each dataset in Figure 4.2. We note that our approach is a generalization of [18] to the multi-agent scenario. While we train only on multi-pedestrian sequences, we can always perform single agent forecasting by setting the social compliance feature to zero (detailed in Section 4.4).

4.1 Metrics

We measured the performance of our forecasting model using two different metrics: (1) negative log loss (NLL) for consistency and (2) state collision rate (SCR) for robustness. Firstly, NLL computes the likelihood of drawing a certain trajectory from learned policy. It is defined as:

$$NLL(\boldsymbol{s}) = -\sum_t \log \pi^{(t)}(\boldsymbol{a}|\boldsymbol{x}), \tag{4.2}$$

where trajectory $\boldsymbol{s}$ denotes a sequence of state $\boldsymbol{x}$, and action $\boldsymbol{a}$ corresponds to the transition between two states. In our case, both $\boldsymbol{s}$ and $\boldsymbol{a}$ are computed from ground truth trajectory sequences.



Town Centre Dataset Zara Dataset LIDAR Trajectory Dataset

Figure 4.2: An overview of the datasets we used in the experiments.

NLL thus describes how consistent the learned policy π is when applied to the demonstrated examples. In addition to NLL, we also employ a metric based on the behavior of multiple people to assess the robustness of a forecasting model. While NLL is a good metric for quantifying how well a model describes the behavior of a single agent, it ignores the interactions among various agents. For instance, walking straight towards goal may suffice for one person due to low NLL, but not when there are multiple people due to potential collisions (Figure 1.1(b)). Thus, we introduce the second metric, state collision rate (SCR). It is defined as:

$$SCR = \frac{1}{Z} \sum_{t} \sum_{n \neq m} D_n^{(t)}(\boldsymbol{x}) D_m^{(t)}(\boldsymbol{x}), \tag{4.3}$$

where n denotes pedestrian ID, $D_n^{(t)}(\boldsymbol{x})$ 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, and Z refers to the normalization factor. By taking into account the distribution of multiple pedestrians and taking the intersection jointly, the expected state visit counts of all agents represent a region of collision.

4.2 Multi-Pedestrian Forecasting Performance

We validate the effectiveness of our model by performing comparative experiments against the following three baselines.

N-Independent MDP (nMDP). This baseline model is the approach of [41] applied to images to forecast the trajectory of a single pedestrian [18]. We extend their approach to meet our multi-agent scenario. In particular, 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 N-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) in advance and incorporate such information into the reward function. The N pedestrians are thus no longer independent as in nMDP.

mTA. Based on the work of Pellegrini *et al.*[27], the third approach is a modified Trajectory Avoidance (mTA) model. In [27] 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 than the original model described in [27] where every

NLL	nMDP[18]	MDPCV	mTA	FP	FP + Speed
Zara [22]	46.5396	46.9549	43.3834	**42.1426**	-
Town Center[4]	14.4797	14.4011	14.2471	12.5804	**10.892**
LIDAR Trajectory	92.5579	93.1747	91.9748	**87.4680**	-

SCR	nMDP[18]	MDPCV	mTA	FP	FP + Speed
Zara [22]	0.144	0.114	0.065	**0.013**	-
Town Center[4]	0.215	0.213	0.120	0.052	**0.049**
LIDAR Trajectory	0.133	0.1050	0.056	**0.009**	-

Table 4.1: Comparative analysis between different approaches.

	Young	Old	Male	Female
Average Speed	1.98	1.25	1.78	1.53

	Age	Gender	Body Direction
Accuracy	82.31%	78.44%	65.60%

Table 4.2: (Left) Average speed (grids/frame) of people with different visual attributes. (Right) Accuracy of the visual attributes classifiers.

agent knows the positions and velocities of others. The information is not available due to our problem setup, *i.e.*, single image input.

For all three datasets, we performed 5-fold cross validation. The forecasting window size is set to $W = 3$ and forecasting period is $\tau = 1$. The results of the proposed approach with comparisons are shown in Table 4.1. 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 only on the Town Center dataset as the resolution of the Zara dataset is too low and the LIDAR-based Trajectory dataset does not provide visual information. We collected ≈ 16000 pedestrian patches from the Town Centre Dataset [4], with three labels for each patch, *i.e.* age, gender, and body orientation. As the TownCentre Dataset already provides the ground truth data for trackers, we do not need to label age and gender for all image patches directly. We ask 2 in-house annotators to label 5 randomly sampled patches for each pedestrian. If the labels for a pedestrian are consistent, we treat them as ground truth and propagate to other frames of a track. Otherwise, we repeat. We also assume that one's body orientation is well aligned with one's trajectory as in [8]. We can thus automatically compute the label for all patches using the ground truth trajectories. Figure 4.3 shows how we generate labels for age and gender. The images are randomly split into ≈ 12000 training images and ≈ 4000 validation images by pedestrians. We train a deep classifier using the network structure mentioned in Section 3.4 and the performance on the validation set is shown in Table 4.2(right). We also computed the average speed of people with different attributes as shown in Table 4.2(left). With the speed statistics and the output of the deep network model, we compute an individualized model for each pedestrian. From Table 4.1 we can observe that our model performs better after considering pedestrian's visual appearance and individualize the predictive model.

Selected qualitative results of predicted trajectories are shown in Figure 4.4. Each pedestrian is marked with a colored bounding box, and a corresponding forecasting distribution is shown in the same color. The more saturated the color is, the higher the probability. Many of the predicted trajectories are smoothly curved (and do not exhibit abrupt changes in trajectories), indicating that the proposed approach mimics human behavior of taking preemptive actions to avoid collisions. We note that we consider *all* pedestrians for quantitative experiments but only visualize forecast distributions for a limited number of pedestrian to improve visualization.

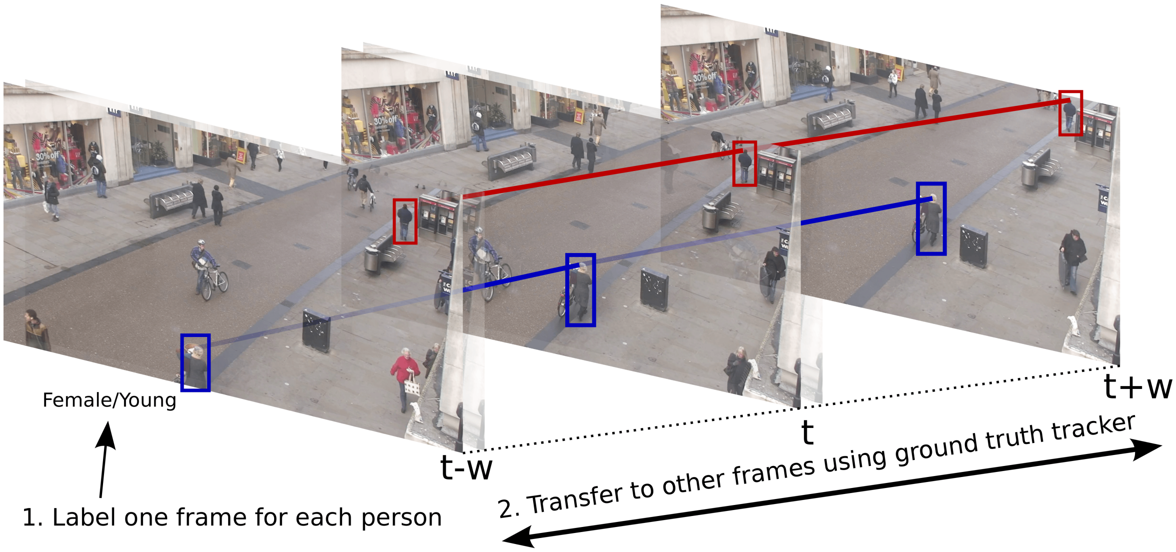

Female/Young
1. Label one frame for each person
t-w
t
t+w
2. Transfer to other frames using ground truth tracker

Figure 4.3: **Labeling attributes for Town Centre dataset:** As the Town Centre dataset already provides the ground truth data for trackers, we do not need to label age and gender for all image patches directly. We simply annotate each pedestrian once and then propagate the labels to other frames of a track. We can also obtain the body orientation label for free by assuming that one's body orientation is well aligned with one's trajectory as in [8].

4.3 Forecasting Window Size and Forecasting Period

We now take a deeper look at the parameters that define the recursive reasoning parameters of the fictitious play algorithm. How far into the future (forecasting window) should the model predict others' movements? How often (forecasting period) should each agent update predictions? In order to better understand how forecasting window size W and forecasting period τ affect the performance of our approach, we perform grid search over the two parameters on the Town Centre dataset, with the estimated speed information included.

From Table 4.3, we can see that given a fixed forecasting window size W, NLL and SCR always increase (gets worse) as the forecasting period τ becomes larger. This phenomenon implies that no matter how far one sees into the future, it is better to predict the actions of others as frequent as possible. This result is reasonable since a larger τ means we assume no changes in the dynamics of the scene over a longer period of time. As for fixing τ and increasing W, SCR decreases initially and then increases with W as expected. This is reasonable since at first the more future steps one forecasts, the more conservative one would be, and therefore less likely to collide with another person. However, as W becomes very large, all agents become overly conservative and may result in collision. Note that NLL does not have a consistent relationship with W, and the breaking point (where scores change from decreasing to increasing) for NLL and SCR is different with various τ. However, in most cases both metrics reach their minimum when $W = 3$. This result to some extent implies that pedestrians in our datasets tend to forecast three steps into the future of others.

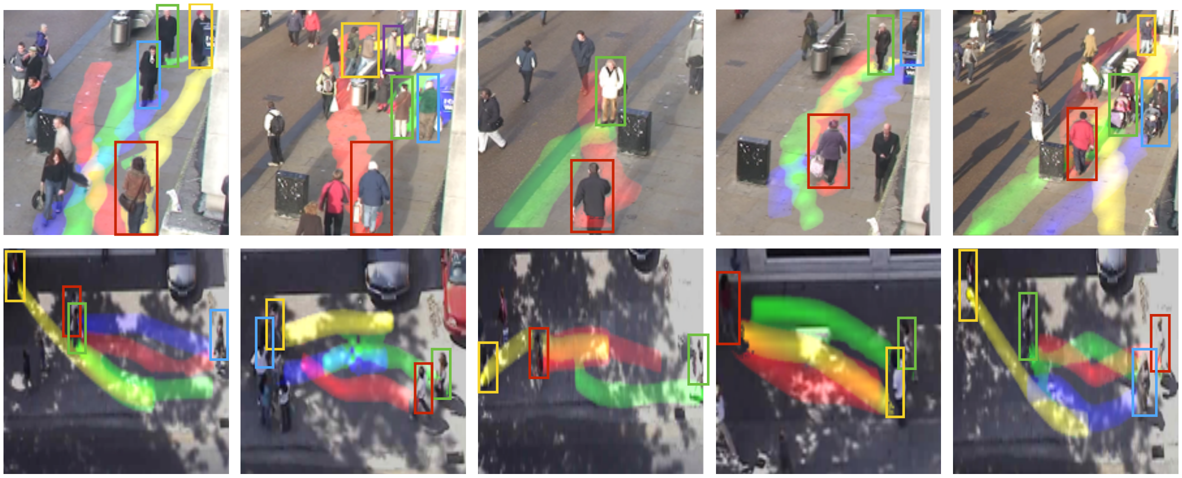


Figure 4.4: Qualitative results show smoothly curving paths indicating predictive collision avoidance.

NLL		Forecasting Period τ			
		1	3	5	7
W	1	14.6947	14.7970	15.7602	15.909
	3	**10.892**	15.5158	15.7619	15.7657
	5	13.3353	15.7070	15.5890	16.1033
	7	14.2277	15.6970	15.7940	16.2070

SCR		Forecasting Period τ			
		1	3	5	7
W	1	0.0735	0.1187	0.167	0.1830
	3	**0.0493**	0.1199	0.1390	0.1606
	5	0.0934	0.1200	0.1287	0.1460
	7	0.1052	0.1221	0.1320	0.1427

Table 4.3: Performance of our model using various fictitious play parameters. W denotes forecasting window (how far into future one should predict), while τ refers to forecasting period (how often one should predict).

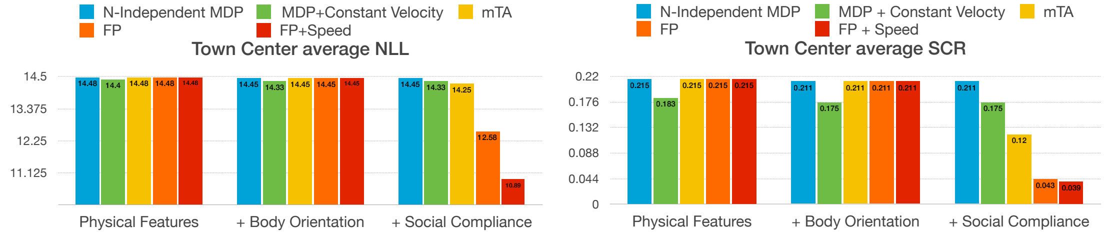


Figure 4.5: Ablative Feature Analysis: Social compliance feature yields strongest predictive power (smaller scores are better).

4.4 Features for Forecasting

We further evaluate the effects of the features used in our proposed model. Three models with different feature sets are trained in advance: (1) models trained *only* on physical features, (2) models trained on both physical and body orientation features $+$ *Body Orientation*, and (3) models trained on all features ($+$ *Social Compliance*). We set forecasting window size $W = 3$ and forecasting period $\tau = 1$ as before.

The average NLL and SCR for the Town Center dataset using different features are shown in Fig. 4.5. The performance of other approaches are also shown for reference. Note that N-Independent MDP and MDP+Constant Velocity still consider only physical features and body orientation features even in the $+$ *Social Compliance* column, since nMDP assumes all entities in the scene are motion-less. Also, our model performs almost identical to nMDP when con-

NLL	nMDP[18]	MDPCV	mTA	FP	FP + Speed
Zara [22]	98.7343	97.6634	92.8271	**88.5693**	-
Town Center[4]	33.8454	33.3213	31.5433	27.5732	**27.2136**

SCR	nMDP[18]	MDPCV	mTA	FP	FP + Speed
Zara [22]	0.186	0.175	0.095	**0.021**	-
Town Center[4]	0.323	0.281	0.170	0.093	**0.066**

Table 4.4: Destination Forecasting Performance (smaller values are better).

sidering only physical 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 (and there is no social compliant feature), our proposed model reduces to nMDP. The most important result is 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 14.2471, resulting in a 23.5% improvement in the NLL.

4.5 Destination Forecasting

In previous sections, we have assumed that the destinations of each pedestrian is known to understand the role of social compliance features and to decouple the effect of uncertain about the goal. In many situations, the final destination of a pedestrian is not known and needs to be inferred. Following [18], we densely generate potential goals on the map and perform the same forecasting experiment, which can be done effectively by inverting the role of goals and start locations. For more details, please refer to [18].

We evaluate the performance of our destination forecasting model on the Town Centre dataset and the Zara dataset. The only information used is the start locations for each pedestrian. No observations are included. Results show that our FP based approach consistently outperforms others even without knowing the destinations ahead of time. The absolute performance of all models degrade due to uncertainty about the goal. From Table 4.4 we observe that mTA still performs the second best, while nMDP still performs the worst. Together with Table 4.1, experimental results show that our model which models the interplay and visual evidence outperforms the baseline models.

Chapter 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.

Bibliography

[1] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2211–2218. IEEE, 2014. 2

[2] Saad Ali and Mubarak Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 2007. 2

[3] Saad Ali and Mubarak Shah. Floor fields for tracking in high density crowd scenes. In *Computer Vision–ECCV 2008*, 2008. 2

[4] Ben Benfold and Ian Reid. Stable multi-target tracking in real-time surveillance video. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3457–3464. IEEE, 2011. 3.1, 4, 4, 4.2, 4.2, 4.5

[5] Steven J Brams. *Game theory and politics*. Courier Corporation, 2011. 3.1

[6] George W Brown. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1):374–376, 1951. 1, 3.1

[7] Brais Cancela, Andres Iglesias, Manuel Ortega, and Manuel G Penedo. Unsupervised trajectory modelling using temporal information via minimal paths. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2553–2560. IEEE, 2014. 2

[8] Isarun Chamveha, Yusuke Sugano, Daisuke Sugimura, Teera Siriteerakul, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Head direction estimation from low resolution images with scene adaptation. *Computer Vision and Image Understanding*, 117(10):1502–1511, 2013. 4.2, 4.3

[9] Haifeng Gong, Jack Sim, Maxim Likhachev, and Jianbo Shi. Multi-hypothesis motion planning for visual object tracking. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 619–626. IEEE, 2011. 2

[10] Edward T Hall. The hidden dimension. 1966. *Garden City*, 1966. 3.3

[11] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 2, 3.3

[12] De-An Huang and Kris M Kitani. Action-reaction: Forecasting the dynamics of human interaction. In *Computer Vision–ECCV 2014*, pages 489–504. Springer, 2014. 2

[13] Ashesh Jain, Hema S Koppula, Bharad Raghavan, Shane Soh, and Ashutosh Saxena. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In

Proceedings of the IEEE International Conference on Computer Vision, pages 3182–3190, 2015. 2

[14] Ashesh Jain, Avi Singh, Hema S Koppula, Shane Soh, and Ashutosh Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. *arXiv preprint arXiv:1509.05016*, 2015. 2

[15] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto. Intent-aware long-term prediction of pedestrian motion. *Proceedings of the International Conference on Robotics and Automation (ICRA)*, May 2016. 2

[16] Kihwan Kim, Matthias Grundmann, Ariel Shamir, Iain Matthews, Jessica Hodgins, and Irfan Essa. Motion fields to predict play evolution in dynamic sport scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 840–847. IEEE, 2010. 2

[17] Sujeong Kim, Stephen J Guy, Wenxi Liu, David Wilkie, Rynson WH Lau, Ming C Lin, and Dinesh Manocha. Brvo: Predicting pedestrian trajectories using velocity-space reasoning. *The International Journal of Robotics Research*, page 0278364914555543, 2014. 2

[18] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. 2012. 2, 3, 3.2, 4, 4, 4.2, 4.5

[19] Henrik Kretzschmar, Markus Kuderer, and Wolfram Burgard. Learning to predict trajectories of cooperatively navigating agents. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 4015–4020. IEEE, 2014. 2

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3.4

[21] Namhoon Lee and Kris M Kitani. Predicting wide receiver trajectories in american football. 2

[22] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer Graphics Forum*, volume 26, pages 655–664. Wiley Online Library, 2007. 3.1, 4, 4, 4.2, 4.5

[23] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 935–942. IEEE, 2009. 2

[24] Roger B Myerson. *Game theory*. Harvard university press, 2013. 3.1

[25] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. *Algorithmic game theory*, volume 1. Cambridge University Press Cambridge, 2007. 3.1

[26] Hyun S Park, Eakta Jain, and Yaser Sheikh. 3d social saliency from head-mounted cameras. In *Advances in Neural Information Processing Systems*, pages 431–439, 2012. 2

[27] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 261–268. IEEE, 2009. 2, 4.2

[28] Matthew Rabin. Incorporating fairness into game theory and economics. *The American economic review*, pages 1281–1302, 1993. 3.1

[29] MS Ryoo, Thomas J Fuchs, Lu Xia, Jake K Aggarwal, and Larry Matthies. Robot-centric activity prediction from first-person videos: What will they do to me'. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 295–302. ACM, 2015. 2

[30] Paul Scovanner and Marshall F Tappen. Learning pedestrian dynamics from the real world. 2

[31] Ozan Sener and Ashutosh Saxena. rcrf: Recursive belief estimation over crfs in rgb-d activity videos. In *Robotics Science and Systems (RSS)*, 2015. 2

[32] Bilge Soran, Ali Farhadi, and Linda Shapiro. Generating notifications for missing actions: Don't forget to turn the lights off! In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4669–4677, 2015. 2

[33] Bulent Tastan and Gita Sukthankar. Leveraging human behavior models to predict paths in indoor environments. *Pervasive and Mobile Computing*, 7(3):319–330, 2011. 2

[34] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian detection aided by deep learning semantic tasks. *arXiv preprint arXiv:1412.0069*, 2014. 3.4

[35] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating the future by watching unlabeled video. *arXiv preprint arXiv:1504.08023*, 2015. 2

[36] Jacob Walker, Abhinav Gupta, and Martial Hebert. Patch to the future: Unsupervised visual prediction. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3302–3309. IEEE, 2014. 2

[37] Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense optical flow prediction from a static image. *arXiv preprint arXiv:1505.00295*, 2015. 2

[38] Dan Xie, Sinisa Todorovic, and Song-Chun Zhu. Inferring" dark matter" and" dark energy" from videos. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2224–2231. IEEE, 2013. 2

[39] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3488–3496, 2015. 2

[40] Xiao-Tong Yuan, Xiaobai Liu, and Shuicheng Yan. Visual classification with multitask joint sparse representation. *Image Processing, IEEE Transactions on*, 21(10):4349–4360, 2012. 3.4

[41] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. 2008. 3.2, 4.2