Learning Social Etiquette: Human Trajectory Understanding in Crowded Scenes

Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, Silvio Savarese

CVGL,
Stanford University
{arobicqu,alahi,amirabs,ssilvio}@stanford.edu

Abstract. Humans navigate crowded spaces such as a university campus by following common sense rules based on social etiquette. In this paper, we argue that in order to enable the design of new target tracking or trajectory forecasting methods that can take full advantage of these rules, we need to have access to better data in the first place. To that end, we contribute a new large-scale dataset that collects videos of various types of targets (not just pedestrians, but also bikers, skateboarders, cars, buses, golf carts) that navigate in a real world outdoor environment such as a university campus. Moreover, we introduce a new characterization that describes the "*social sensitivity*" at which two targets interact. We use this characterization to define "*navigation styles*" and improve both forecasting models and state-of-the-art multi-target tracking - whereby the learnt forecasting models help the data association step.

Keywords: Trajectory Forecasting, Multi-target Tracking, Social Forces, UAV

1 Introduction

When pedestrians or bicyclists navigate their way through crowded spaces such as a university campus, a shopping mall or the sidewalks of a busy street, they follow common sense conventions based on social etiquette. For instance, they would yield the right-of-way at an intersection as a bike approaches very quickly from the side, avoid walking on flowers, and respect personal distance. By constantly observing the environment and navigating through it, humans have learnt the way other humans typically interact with the physical space as well as with the targets that populate such spaces *e.g.*, humans, bikes, skaters, electric carts, cars, toddlers, etc. They use these learned principles to operate in very complex scenes with extraordinary proficiency.

Researchers have demonstrated that it is indeed possible to model the interaction between humans and their surroundings to improve or solve numerous computer vision tasks: for instance, to make pedestrian tracking more robust and accurate [1,2,3,4,5], to enable the understanding of activities performed by groups of individuals [6,7,8,9], to enable accurate prediction of target trajectories

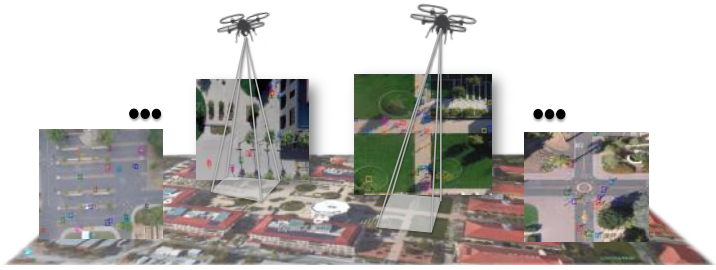

Fig. 1. We aim to understand human social navigation in a multi-class setting where pedestrians, bicyclists, skateboarders and carts (to name a few) share the same space. To that end, we have collected a new dataset with a quadcopter flying over more than 100 different crowded campus scenes.

in future instants [10,11,12,13]. Most of the time, however, these approaches operate under restrictive assumptions whereby the type and number of interactions are limited or the testing environment is often contrived or artificial.

In this paper, we argue that in order to learn and use models that allow mimicking, for instance, the remarkable human capability to navigate in complex and crowded scenes, the research community needs to have access to better data in the first place. To that end, we contribute a new large scale dataset that collects videos of various types of targets (not just pedestrians, but also bikes, skateboarders, cars, buses, golf carts) that navigate in a real world outdoor environment such as a university campus. Our dataset comprises of more than 100 different top-view scenes for a total of 20,000 targets engaged in various types of interactions. Target trajectories along with their target IDs are annotated which makes this an ideal testbed for learning and evaluating models for multi-target tracking, activity understanding and trajectory prediction at scale (see Figure 1 and 2).

Among all the problems discussed above, in this paper we are interested in evaluating techniques related to two classes of problems: i) target trajectory forecasting - whereby the ability to comply to social etiquettes and common sense behavior is critical, ii) Multi-Target Tracking (MTT) - whereby the learnt forecasting model is used to enhance tracking results. In particular, we believe that our new dataset creates the opportunity to generalize state-of-the-art methods for understanding human trajectory, and evaluate them on a more effective playground. For instance, two leading families of methods for target trajectory forecasting (Social Forces [14,15,1,2] and Gaussian Processes [16,17,12]) have shown promising results on existing datasets [18,11]; however, they have never been tested at scale and in real-world scenarios where multiple classes of targets are present (i.e., not just pedestrian but also cars, bikes, etc.) as part of a complex ecosystem of interacting targets.

In addition to evaluating state-of-the-art forecasting and tracking methods, in this paper we also introduce a novel characterization that describes the "*so-*

cial sensitivity" at which two targets interact. It captures both the preferred distance a target wants to preserve with respect to its surrounding as well as when (s)he decides to avoid other targets. Low values for the social sensitivity feature means that a target motion is not affected by other targets that are potentially interacting with it. High values for the social sensitivity feature means that the target navigation is highly dependent on the position of other targets. This characterization allows to define the "*navigation style*" targets follow in interacting with their surrounding. We obtain different classes of navigation styles by clustering trajectory samples in the *social sensitivity space* (see Figure 3 for examples). This allows to increase the flexibility in characterizing various modalities of interactions - for instance, some pedestrians may look more aggressive while walking because they are in rush whereas others might show a milder behavior because they are just enjoying their walk. Navigation style classes are used to select the appropriate forecasting model to best predict targets' trajectories as well as improve multi-target tracking. We believe that the ability to model social sensitivity is a key step toward learning common sense conventions based on social etiquette for enhancing forecasting and tracking tasks.

We present an extensive experimental evaluation that compares various state-of-the-art methods on the newly proposed dataset, and demonstrates that our social sensitivity feature and the use of navigation style enable better prediction and tracking results than previous methods that assume that all the targets belong to the same class (*i.e.*, follow the same navigation style).

2 Previous Work

A large variety of methods has been proposed in the literature to describe, model and predict human behaviors in a crowded space. Here we summarize the most relevant methods for human trajectory forecasting and multi-target tracking.

Human trajectory forecasting. An exhaustive study of crowd analysis is introduced by Treuille *et. al.* [19]. Antonini *et. al.* use the Discrete Choice Model to synthesize human trajectories in crowded scenes [20,21]. Other methods [22,17,12] use Gaussian Processes to forecast human trajectories. They avoid the problems associated with discretization and their generated motion paths are smooth. Unfortunately, they often assume that the location of the destination is known. More recently, a set of methods use Inverse Reinforcement Learning [10,23,24] whereby a reward (or cost) function is learnt that best explains the final decisions [25]. While these techniques have shown to work extremely well in several applications [26,25,27], they assume that all feature values are known and static during each demonstrated planning cycle. They have been used to mainly model human and static space interaction as opposed to the dynamic content.

The most popular method for multi-target trajectory forecasting remains the *Social forces* (SF) model by D. Hellbing and P. Molnar [15]. Targets react to energy potentials caused by the interactions with other targets and static obstacles through forces (repulsion or attraction). The SF model has been extensively

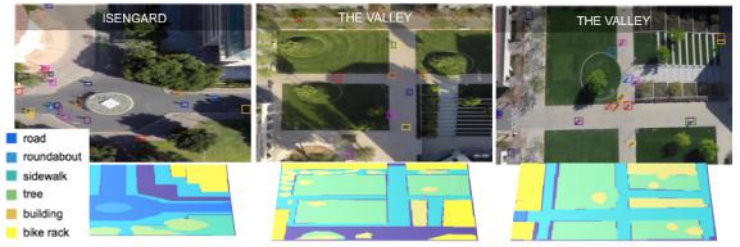

Fig. 2. Some examples of the scenes captured in our dataset. We have annotated all the targets (with bounding boxes) as well as the static scene semantics. The color codes associated to target bounding boxes represents different track IDs.

used in robotics [28], and in the context of target tracking [1,2,29,30,31,32,33]. All these previous work use a single set of parameters to model multiple targets. We argue and show in the remainder of this paper that a single set of parameters is too limited to model all the navigation styles in complex crowded scenes when multiple classes of targets are present (pedestrians, bikers, skateboarders,...).

Multi-Target Tracking. Over the past decade, Multi-Target Tracking (MTT) algorithms have made great progress in solving the data association problem as a graph theoretic problem [34,35,36,30,37,38]. Several methods have incorporated the Social Forces (SF) model to improve the motion prior [1,2,3,4,5]. Recently, Xiang *et. al.* [39] demonstrate the power of a strong appearance model over all these previous work. They reached state-of-the-art performance over the publicly available MTT challenge [40]. In this work, we use their method and demonstrates the impact of our "social sensitivity" feature in crowded multi-class complex scenes.

In the next sections, we first present our collected dataset. Then, we introduce our social sensitivity feature. In Section 5, we share details behind our forecasting and tracking model. Finally, we conclude with a detailed evaluation of our forecasting task, and its impact on the Multi-Target Tracking task.

3 Campus Dataset

We aim to learn the remarkable human capability to navigate in complex and crowded scenes. Existing datasets mainly capture the behavior of humans in spaces occupied by a single class of target, *e.g.*, pedestrian-only scenes [18,11,31]. However, in practice, pedestrians share the spaces with other classes of targets such as bicyclists, or skateboarders to name a few. For instance, on university campuses, a large variety of these targets interacts at peak hours. We want to study social navigation in these complex and crowded scenes occupied by several classes of targets. Datasets such as [41,42] do contain multiple classes of objects but are either limited in the number of scenes (just one for [41]), or in the number of classes of moving targets (just pedestrians in [42]).

To the best of our knowledge, we have collected the first large-scale dataset that has images and videos of various classes of targets that are moving and interacting in a real-world university campus. The dataset comprises more than 19K targets consisting of 11.2K pedestrians, 6.4K bicyclists, 1.3k cars, 0.3K skateboarders, 0.2K golf carts, and 0.1K buses. Although only videos of campus scenes are collected, the data is general enough to capture all type of interactions:

- target-target interactions, *e.g.*, a bicyclist avoiding a pedestrian,
- target-space interactions, *e.g.*, a skateboarder turning around a roundabout.

Target-target interactions We say that two targets interact when their collision energy (described by Equation 1) is non-zero, *e.g.*, a pedestrian avoiding a skateboarder. These interactions involve multiple physical classes of targets (pedestrians, bicyclists, or skateboarders to name a few), resulting into 185K annotated target-target interactions. We intentionally collected data at peak hours (between class breaks in our case) to observe high density crowds. For instance, during a period of 20 seconds, we observe in average from 20 to 60 targets in a scene (of approximately $900m^2$).

Target-space interactions. We say that a target interacts with the space when its trajectory deviates from a linear one in the absence of other targets in its surrounding, *e.g.*, a skateboarder turning around a roundabout. To further analyze these interactions, we also labeled the scene semantics of more than 100 static scenes with the following labels: road, roundabout, sidewalk, grass, building, and bike rack (see Figure 2). We have approximately 40k "target-space" interactions. In our model, the whole target space interaction is implicitly considered in the Social Force model. We only take dynamic obstacles into account. However, in most common scenes, people will also try to avoid static obstacles. Similar to [18] we model such obstacles as agents with zero velocity.

Dataset	Frames	Targets	Interactions	Bi	Ped	Skate	Carts	Car	Bus
Isengard	134079	2044	6472	1004	926	57	19	23	15
Hobbiton	138513	3821	14084	163	2493	24	18	1065	58
Edoras	47864	1186	4684	224	956	2	2	2	0
Mordor	139364	4542	68459	2594	1492	111	154	165	26
Fangorn	249967	3126	45520	1017	1991	50	30	27	11
Valley	219712	4845	46062	1362	3358	89	21	10	5
Total	929499	19564	185281	6364	11216	333	244	1292	115

Table 1. Our campus dataset characteristics. We group the scenes and refer to them using fictional places from the "Lord of the Rings". Bi = bicyclist, Ped = pedestrian, Skate = skateboarders.

Tables 1 presents more details on our collected dataset. The scenes are grouped into 6 areas based on their physical proximity on campus. Each scene is captured with a 4k camera mounted on a quadcopter platform (a 3DR solo)

hovering above various intersections on a University campus at an altitude of approximately eighty meters. The videos have a resolution of 1400x1904 and have been processed (*i.e.* undistorted and stabilized). Targets are annotated with their class label and their trajectory in time and space is identified (see supplementary material for more details).

4 Modeling Social Sensitivity

We claim that modeling human trajectory with a single navigation style is not suitable for capturing the variety of social behaviors that targets exhibit when interacting in complex scenes. We believe that conditioning such models on *navigation style* (*i.e.*, the way targets avoid each other) is a better idea and propose a characterization (feature) which we call *social sensitivity*. Given this characterization, we hence assign a navigation style to each target to better forecast its trajectory and improve tracking.

Social Sensitivity feature. Inspired by the Social Forces model (SF) [1], we model targets' interactions with an energy potential E_{ss}. A high potential means that the target is highly sensitive to others. We define E_{ss} as follows:

At each time step t, the target i is defined by a state variable $s_i^{(t)} = \{\mathbf{p}_i^{(t)}, \mathbf{v}_i^{(t)}\}$, where $\mathbf{p}_i^{(t)}$ is the position, and $\mathbf{v}_i^{(t)}$ the velocity. The energy potential encoding the social sensitivity is computed as follows:

$$E_{ss}(\mathbf{v_i^{(t)}}; s_i, \mathbf{s_{-i}} | \sigma_d, \sigma_w, \beta) = \sum_{j \neq i} w(s_i, s_j) \exp\left(-\frac{d^2(\mathbf{v}, s_i, s_j)}{2\sigma_d^2}\right), \tag{1}$$

with $w(s_i, s_j)$ defined as:

$$w(s_i, s_j) = \exp\left(-\frac{|\Delta\mathbf{p}_{ij}|}{2\sigma_\omega}\right) . \left(\frac{1}{2}\left(1 - \frac{\Delta\mathbf{p}_{ij}}{|\Delta\mathbf{p}_{ij}|}\frac{\mathbf{v}_i}{|\mathbf{v}_i|},\right)\right)^\beta, \tag{2}$$

and

$$d^2(\mathbf{v}, s_i, s_j) = \left|\Delta\mathbf{p}_{ij} - \frac{\Delta\mathbf{p}_{ij}(\mathbf{v} - \mathbf{v}_j)}{|\mathbf{v} - \mathbf{v}_j|^2}(\mathbf{v} - \mathbf{v}_j)\right|. \tag{3}$$

The energy E_{ss} is modeled as a product of Gaussians where the variances $\sigma_{w,d}$ represent the distances at which other targets will influence each other. For instance, if two targets i, j are close to each other ($\Delta\mathbf{p}_{ij}$ is small), E_{ss} will be large when $\sigma_{w,d}$ are small.

We define the parameter $\Theta_{ss} = \{\sigma_d, \sigma_w, \beta\}$ as the social sensitivity feature and interpret its dimension as follows:

- σ_d is the preferred distance a target maintains to avoid collision,
- σ_w is the distance at which a target reacts to prevent a collision (distance at which (s)he starts deviating from its linear trajectory),

– and β controls the peakiness of the weighting function.

In other words, the parameters $\{\sigma_d, \sigma_w, \beta\}$ aim at describing how targets avoid each others - i.e., their social sensitivity. We now present how we infer the parameters Θ_{ss} at training and testing time.

Training. At training time, since we observe all targets' velocities, V^{train}, we could learn a unique set of parameters, *i.e.*, a single value for social sensitivity, that minimizes the energy potential as follows (similarly to what previous methods do [1,2,3,4,5]):

$$\{\sigma_d, \sigma_w, \beta\} = \underset{\{\sigma_d, \sigma_w, \beta\}}{\operatorname{argmin}} \left(\sum_{i=1}^{T-1} E_{ss}(v_i^{train}, s_i, s_{-i} | \sigma_d, \sigma_w, \beta) \right), \tag{4}$$

where T is the number of targets in the training data. This minimization is operated with an interior-point method and is set with the following constraint on σ_d: $\sigma_d > 0.1$ (it specifies that every target can't have a "vital space" smaller than 10cm).

As mentioned previously, however, we claim that learning a unique set of parameters is not suitable when one needs to deal with complex multi-class target scenarios whereby targets can have different social sensitivity. To validate this claim, we plot in Figure 3 each target into a *social sensitivity space* where the x-axis is the σ_d values and the y-axis is the σ_w ones. These data are computed using training images from our dataset (see Section 6 for more details). We did not plot the third parameter β since it does not change much across targets. Even if our approach can handle an arbitrary number of classes, we cluster the points into four clusters for illustration purposes. Each cluster corresponds to what we define as a "navigation style". A navigation style describes the sensitivity of a target to its surrounding. We illustrate on the sides of Figure 3 how targets follow different strategies in avoiding each other as different navigation styles are used.

Thanks to the above analysis of the *social sensitivity space*, at training, we solve Equation 4 for each target to get its social sensitivity feature. We then cluster the points with K-mean clustering to have N number of clusters. Each cluster represents a navigation style. In Section 6, we study the impact of the number of clusters used by our method on the forecasting accuracy in Table 4.

Testing. At test time, we observe the targets until time t, and want to assign a navigation style.

In the presence of other targets, we solve Equation 5 for each specific target i at time t:

$$\{\sigma_d(i), \sigma_w(i), \beta(i)\} = \underset{\{\sigma_d(i), \sigma_w(i), \beta(i)\}}{\operatorname{argmin}} \left(E_{ss}(v_i^t, s_i, s_{-i} | \sigma_d(i), \sigma_w(i), \beta(i)) \right). \tag{5}$$

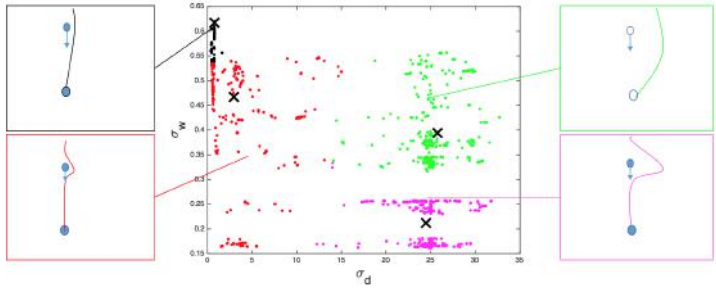

Fig. 3. Illustration of the social sensitivity space where we have illustrated how targets avoid each other with four navigation styles (from a top view). Each point in the middle plot is a target. The x-axis is the preferred distance σ_d a target keeps with its surrounding targets, and y-axis is the distance σ_w at which a target reacts to prevent a collision. Each color code represents a cluster (a navigation style). Even if our approach can handle an arbitrary number of classes, we only use 4 clusters for illustration purposes. In this plot, the green cluster represents targets with a mild behavior, willing to avoid other targets as much as possible and considering them from afar, whereas the red cluster describes targets with a more aggressive behavior and with a very small safety distance, considering others at the last moment. We illustrate on the sides of the plot examples of how targets follow different strategies in avoiding each other as different navigation styles are used.

We obtain the social sensitivity feature $\Theta_{ss}(i) = \{\sigma_d(i), \sigma_w(i), \beta(i)\}$ for each target i. Given the clusters found at training, we assign each $\Theta_{ss}(i)$ to its corresponding cluster, i.e., navigation style.

In the absence of interactions, a target takes either a "neutral" navigation style (when entering a scene) or inherit the last inferred class from the previous interaction. The "neutral" navigation style is the most popular one (in green in Figure 3). In figure 4, we show that when the target is surrounded by other targets, its class changes with respect to its social sensitivity.

5 Forecasting and Tracking with Social Sensitivity

Our new collected dataset creates the opportunity to study methods for trajectory forecasting and multi-target tracking, and evaluate them on a large-scale broad setting, *i.e.* a space occupied by several classes of targets. Thanks to our proposed social sensitivity feature, we have more flexibility in modeling target interactions to forecast future trajectories. In the remaining of this section, we present the details behind our forecasting model driven by social sensitivity.

Fig. 4. Illustration of the class assignment for each target. The same color represents the same navigation style (cluster) described in Figure 3. Note that for a given target its class changes across time regardless of its physical class (*i.e.*, whether it is a pedestrian, bike, etc.). When the target is surrounded by other targets, its class changes with respect to its social sensitivity. In this scene, first we can observe a cyclist (shown as label 1 in the images) belonging to a black cluster, *i.e.*, being aggressive in his moves, then belonging to some milder clusters (purple and green). We also can see the evolution of a group of pedestrians (shown as labels 2,3) in the images), initially "mild" (green at $T = 1$), who become red at time $T = 3$ at which they decide to overtake another group and accelerate.

Then, in Section 5.2, we show how to use our forcasting model on multi-target tracking.

5.1 Forecasting multiple classes of targets

Problem formulation. Given the observed trajectories of several targets at time t, we aim to forecast their future positions over the next N time frames (where N is in seconds).

We adapt the Social Forces model [1] from single class to multiple classes. Each target makes a decision on its velocity $\mathbf{v}_i^{(t+1)}$. The energy function, E_Θ, associated to every single target is defined as:

$$\begin{aligned} E_\Theta(\mathbf{v^{t+1}}; s_i, \mathbf{s_{-i}}) = \lambda_0(c)E_{damp}(\mathbf{v^{t+1}}; s_i) + \lambda_1(c)E_{speed}(\mathbf{v^{t+1}}; s_i) \\ +\lambda_2(c)E_{dir}(\mathbf{v^{t+1}}; s_i) + \lambda_3(c)E_{att}(\mathbf{v^{t+1}}; s_i) + \lambda_4(c)E_{group}(\mathbf{v^{t+1}}; s_i, \mathbf{s}_{A_i}) \\ +E_{ss}(\mathbf{v^{t+1}}; s_i, \mathbf{s_{-i}} | \sigma_d(v^t), \sigma_w(v^t), \beta) \end{aligned} \tag{6}$$

where $\Theta = \{\lambda_0(c), \lambda_1(c), \lambda_2(c), \lambda_3(c), \lambda_4(c), \sigma_d(v^t), \sigma_w(v^t), \beta\}$ and c is the navigation class. More details on the definition of each of the energy terms can be found in [1].

In our work, we propose to compute σ_d, and σ_w directly from the observed velocity v^t using Equation 5. Both distances σ_d, and σ_w will then be used to identify the navigation class c. For each class c, the parameter Θ can be learned from training data by minimizing the energy in Equation 6.

Time Complexity At test time, we only need to infer 3 parameters instead of few dozen at training time. Once these 3 parameters are inferred, we use the result from our k-means clustering to get the remaining parameters. Consequently, the

computation cost went from 1 min (to infer all parameters) to 0.1 sec (to infer three parameters) (per frame and agent with a matlab implementation).
There is an additional computational complexity of $\mathcal{O}(nkdi)$ for k-means which comes at negligible computational cost (less than 1 ms), where n is the number of d-dimensional vectors (in this application 2), k the number of clusters (number of behavioral classes) and i the number of iterations needed until convergence which is not more than 10 iterations.

5.2 Multi-target Tracking

Problem formulation. Given the detected targets at each time frame (using for instance a target detector [43], or a background subtraction method [44]), we want to link the detection results across time to form trajectories, commonly referred to as tracking-by-detection.

As mentioned in Section 2, we modify the Multi-target Tracking (MTT) algorithm from Xiang *et. al.* [39] to utilize our multi-class forecasting model based on social sensitivity. They formulate the MTT problem as a Markov Decision Process (MDP), which seeks to model the trajectory of targets according to a set of valid states (*e.g.*, $s_{tracked}, s_{lost}$) and transitions. They construct an approach to data association by computing a feature vector ϕ_i^t that describes the appearance of the targets in each of these possible states. They furthermore use a linear motion prior to reason on the navigation of targets, to thus determine a heuristic as to where a target should generally lie in future frames.

In order to evaluate the effectiveness of social sensitivity, we replace their linear motion prior with our multi-class forecasting method. More specifically, we modify ϕ_i^t, the feature vector for target i at time t as follows: Given the coordinates x_i^t, y_i^t of the target, we first apply our social force model to obtain a prediction x_i^{t+1}, y_i^{t+1} of the target at the next timestep. Then, given a list of candidate detections D^{t+1} for data association, we compute a normalized Euclidean distances $\{d_1, d_2, \ldots\}$ between each detection and the predicted coordinates, and append e^{-d_j} to ϕ_i^t, where d_j is the distance to detection j. In Section 6, we show the gain in performance from applying this method to our dataset.

6 Experiments

We run two sets of experiments: First, we study the performance of our method on trajectory forecasting problem. Then, we demonstrate the effectiveness of our proposed social sensitivity feature on state-of-the-art multi-target tracking - whereby the learnt forecasting models help the data association step.

6.1 Forecasting accuracy

Datasets and metrics. We evaluate our multi-class forecasting framework on our new collected dataset as well as previous existing pedestrian-only ones [18,11].

Methods	Lin			LTA			SF [1]			IGP [45]			Our SF-mc		
ETH	0.80	0.95	1.31	0.54	0.70	0.77	0.41	0.49	0.59	**0.20**	**0.39**	**0.43**	0.41	0.46	0.59
HOTEL	0.39	0.55	0.63	0.38	0.49	0.64	0.25	0.38	0.37	**0.24**	0.34	0.37	**0.24**	**0.32**	**0.37**
ZARA 1	0.47	0.56	0.89	0.37	0.39	0.66	0.40	**0.41**	0.60	0.39	0.54	**0.39**	**0.35**	**0.41**	0.60
ZARA 2	0.45	0.44	0.91	0.40	0.41	0.72	0.40	0.40	0.68	0.41	0.43	0.42	**0.39**	**0.39**	0.67
UCY	0.57	0.62	1.14	0.51	0.57	0.95	0.48	0.54	0.78	0.61	0.62	1.82	**0.45**	**0.51**	**0.76**
AVERAGE	0.54	0.62	0.97	0.44	0.51	0.75	0.39	0.44	**0.60**	**0.37**	0.46	0.69	**0.37**	**0.42**	**0.60**

Table 2. Pedestrian Only dataset - Our 3 main evaluation methods, ordered as: Mean Average Displacement on all trajectories | Mean Average Displacement on collisions avoidance | Average displacement of the predicted final position (after 4.8 seconds).

Methods	Lin			SF			IGP [45]			SF-Physical			Our SF-mc		
ISENGARD	1.69	1.00	2.84	1.60	0.99	2.32	1.57	1.14	2.64	1.56	0.86	1.83	**1.53**	**0.84**	**1.81**
HOBBITON	1.17	1.01	1.81	**1.11**	0.82	1.70	**1.11**	**0.81**	2.25	1.12	**0.81**	**1.70**	1.12	0.83	**1.70**
EDORAS	0,91	0.83	1.03	0.80	**0.81**	0.89	1.33	0.85	2.61	0.79	**0.81**	**0.89**	**0.78**	0.82	**0.89**
MORDOR	1.72	1.10	3.80	1.38	0.89	2.30	**0.95**	0.69	**1.78**	1.37	0.65	2.30	1.37	**0.60**	2.30
FANGORN	1.02	0.75	2.00	0.94	0.41	1.66	0.96	0.69	1.67	0.90	0.40	**1.51**	**0.89**	**0.36**	**1.51**
VALLEY	1.38	0.86	2.45	1.29	0.87	2.02	1.20	0.75	2.46	1.01	**0.65**	**1.65**	**0.99**	0.66	**1.65**
AVERAGE	1.32	0.93	2.32	1.29	0.79	1.82	1.19	0.82	2.24	1.14	0.70	1.65	**1.11**	**0.69**	**1.64**

Table 3. Campus Dataset - Our 3 main evaluation methods, ordered as: Mean Average Displacement on all trajectories | Mean Average Displacement on collisions avoidance | Average displacement of the predicted final position (after 4.8 seconds).

Our dataset has two orders of magnitude more targets than the combined pedestrian-only datasets. We evaluate the performance of forecasting methods with the following measures: average prediction error over (i) the full estimated trajectory, (ii) the final estimated point, and (iii) the average displacement during collision avoidance's. Similar to [18,11], we observe trajectories for 2.4 seconds and predict for 4.8 seconds. We sub-sample a trajectory every 0.4 second. We also focus our evaluation when non-linear behaviors occur in the trajectories to not be affected by statistically long linear behaviors.

Quantitative and qualitative results. We evaluate our proposed multi-class forecasting framework against the following baselines: (i) single class forecasting methods such as SF [1] and IGP [45], (ii) physical class based forecasting (SF-pc), *i.e.*, using the ground truth physical class, and (iii) our proposed method inferring navigation style of the targets referred to as SF-mc. We present our quantitative results in Tables 2 and 3:

On pedestrian-only dataset (Table 2), our SF-mc performs the same as the single class Social Forces model in ETH dataset, and outperforms other methods in UCY datasets. This result can be justified by the fact that the UCY dataset is considerably more crowded, with more collisions, and therefore presenting different types of behaviors. Non-linear behaviors such as people stopping and talking to each other, walking faster, or turning around each others are more common in UCY than in ETH. Our forecasting model is able to infer these navigation patterns hence better predict the trajectories of pedestrians. We also

	1 [1]	2	4	7	12	18
Mean error	1.14	1.16	1.15	**1.11**	1.12	1.20
Collision error	0.72	**0.68**	0.69	0.69	0.73	0.75
Final position error	1.84	1.74	1.70	**1.64**	1.69	1.80

Table 4. Forecasting error with respect to the number of clusters in our new campus dataset.

report the performance of the IGP model on these pedestrian-only datasets for completeness. While IGP performs better on the less crowded dataset, it does not do well on the crowded ones. Notice that IGP uses the destination and time of arrival as additional inputs (which our method don't use).

On our multi-class dataset (Table 3), we can see that our approach is more accurate on every scenes when a large amount of different classes are present. Our highest gain in performance is visible on the last three scenes, rich in classes and collisions (see Table 1). In HOBBITON and EDORAS scenes, our algorithm, trained on a multi-class dataset, matches the single class Social Forces. This happens because the social sensitivity feature stays the same across targets. In a scene with less number of classes, this could become a drawback, but yet our algorithm can perform with the same accuracy.

In Section 5.1, we present our method to forecast multiple classes of targets where we use the learned navigation styles as classes. One can argue that instead of using the navigation styles, we could use target's class (*e.g.* pedestrian, bicyclist, etc.). Table 3 compares the performance of using navigation style against targets' class (e.g. one parameter per pedestrian, bicyclist, and so on...), referred to as SF-Physical. We use the ground truth class label to associate each target to their corresponding physical class - this gives an upper bound accuracy. Interestingly, both multi-class strategies perform almost the same although our method does not require ground truth physical class labels as it automatically assign the navigation style class to each target as described in Section 5.1.

We further study the impact of the number of navigation styles (clusters) used by our method on the forecasting accuracy in Table 4. The optimal performance is obtained with 7 navigation styles which coincidentally, is very similar to the number of target's class (6 in our dataset). All experiments results in table 3 are given considering 7 clusters.

Once a target is associated to one of the navigation styles, the corresponding parameter θ from Equation 6 is used to predict the trajectory of the target. We can visualize the impact of the navigation style on the prediction. In figure 5, we show the predicted trajectories when several navigation styles are used to perform the forecasting. This shows the need to assign targets into specific classes.

Finally, in figure 6, we show more examples of our predicted trajectories and compare them with previous works. Our proposed multi-class framework outperforms previous methods in crowded scenes. However, in the absence of interactions, all methods perform the same.

	Rcll	Prcn	MT	ML	MOTA	MOTP	MOTAL
MDP [39] + Lin	74.1	80.1	44.18%	**20.9%**	51.5	74.2	55.4
MDP [39] + SF [1]	84.4	91.5	58.13%	25.5%	73.5	77.1	76.3
MDP [39] + our SF-mc	**86.1**	**92.6**	**60%**	23.2%	**75.6**	**78.2**	**79.3**

Table 5. MTT tracking results.

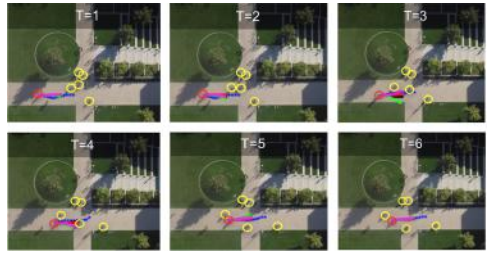

Fig. 5. We show the predicted trajectory of a given target (red circle) in which four different navigation styles are used to perform the prediction. The corresponding predicted trajectories are overlaid on one other and shown with different color codes (the same as those used for depicting the clusters in figure 3). The ground truth is represented in blue. Predicted trajectories are shown for 6 subsequent frames indicated by $T = 1, ..., 6$ respectively. Interestingly, when the target is far away from other targets (no interactions are taking place) the predicted trajectories are very similar to each other (they almost overlap and show a linear trajectory). However, when the red target gets closers to other targets (e.g. the ones indicated in yellow), the predicted trajectories start showing different behaviors depending on the navigation style: a conservative navigation style activates trajectories' prediction that keep large distances to the yellow targets in order to avoid them (green trajectory) whereas an aggressive navigation style activates trajectories' prediction that are not too distant from the yellow targets (red trajectory). Notice that our approach is capable to automatically associate the target to one of the 4 clusters based on the characteristics in the social sensitivity space that have been observed until present. In this example, our approach selects the red trajectory which is the closest to the ground truth's predicted trajectory (in blue).

6.2 Multi-target tracking evaluation

Dataset and metrics . We evaluate the impact of our social sensitivity feature on multi-target tracking using our newly collected dataset which contain images from crossing roads, sidewalks, and many other types of scene semantics with roughly 30 people observed per frame. We use the same evaluation metric as the MTT challenge [40], such as the multi object tracking accuracy (MOTA), or mostly tracked (MT) objects. In details the multiple object tracking accuracy (MOTA) takes into account false positives, missed targets and identity switches, multiple object tracking precision (MOTP) is simply the average distance between true and estimated targets. The other metrics such as mostly tracked (MT) and mostly lost (ML) counts the number of mostly tracked trajectories

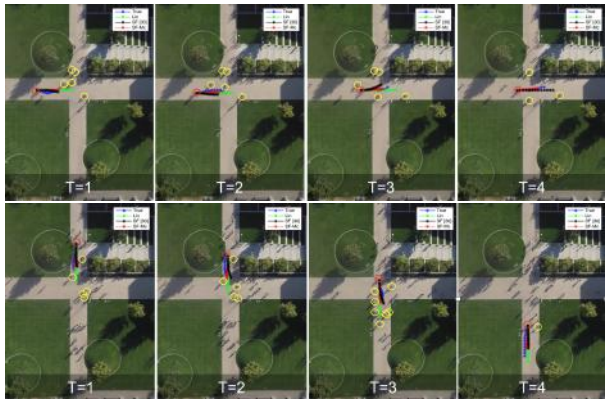

Fig. 6. Illustration of the predicted trajectories by our SF-mc method (in red) across time. Predicted trajectories are shown for 4 subsequent frames indicated by $T = 1, ..., 4$ respectively. We compare them with previous work [1]. The ground truth is represented in blue. Our proposed multi-class framework outperforms previous methods when targets start interacting with other target (t=2,3,4). However, in the absence of interactions (t=1), all methods perform the same.

(more than 80% of the frames) and mostly lost (was not able to track more than 20% of the frames). The full list of metrics can be found in [40].

Quantitative results. We evaluate our proposed MTT algorithm against the following baselines: (i) Xiang's MDP algorithm [39] with a linear motion prior, (ii) [39] with single class forecasting model [1], (iii) [39] with our proposed multi-class forecasting model based on social sensitivity. We show that using our proposed MTT with social sensitivity feature outperforms previous work. Our quantitative results are shown in Table 5.

7 Conclusions

We have presented our efforts to study human navigation at a new scale. We have contributed the first large-scale dataset of aerial videos from multiple classes of targets interacting in complex outdoor spaces. We have presented our work on predicting the trajectories of several classes of targets without explicitly solving the target classification task. We further demonstrate the impact of our forecasting model on multi-target tracking. Future work will study other forecasting methods such as Long Short-Term Memory (LSTM) to jointly solve the prediction task. Finally, by sharing our dataset, we hope that researchers will push the limits of existing methods in modeling human interactions, learning scene specific human motion, or detecting and tracking tiny targets from UAV data.

Acknowledgments We thank Panasonic (1192707-1-GWMSX), ONR (1183788-1-TDZBP), and Ambarella for supporting this project.

References

1. Yamaguchi, K., Berg, A.C., Ortiz, L.E., Berg, T.L.: Who are you with and where are you going? In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1345–1352
2. Pellegrini, S., Ess, A., Van Gool, L.: Improving data association by joint modeling of pedestrian trajectories and groupings. In: Computer Vision–ECCV 2010. Springer (2010) 452–465
3. Leal-Taixé, L., Fenzi, M., Kuznetsova, A., Rosenhahn, B., Savarese, S.: Learning an image-based motion context for multiple people tracking. In: CVPR, IEEE (2014) 3542–3549
4. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: Computer Vision–ECCV 2012. Springer (2012) 215–230
5. Smeulders, A.W., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: an experimental survey. Pattern Analysis and Machine Intelligence, IEEE Transactions on **36**(7) (2014) 1442–1468
6. Xie, D., Todorovic, S., Zhu, S.C.: Inferring" dark matter" and" dark energy" from videos. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 2224–2231
7. Choi, W., Savarese, S.: Understanding collective activitiesof people from videos. Pattern Analysis and Machine Intelligence, IEEE Transactions on **36**(6) (2014) 1242–1257
8. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: Advances in neural information processing systems. (2010) 1216–1224
9. Choi, W., Shahid, K., Savarese, S.: What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, IEEE (2009) 1282–1289
10. Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M.: Activity forecasting. In: Computer Vision–ECCV 2012. Springer (2012) 201–214
11. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. In: Computer Graphics Forum. Volume 26., Wiley Online Library (2007) 655–664
12. Trautman, P., Ma, J., Murray, R.M., Krause, A.: Robot navigation in dense human crowds: the case for cooperation. In: Robotics and Automation (ICRA), 2013 IEEE International Conference on, IEEE (2013) 2153–2160
13. Cucchiara, R., Grana, C., Tardini, G., Vezzani, R.: Probabilistic people tracking for occlusion handling. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Volume 1., IEEE (2004) 132–135
14. Hughes, R.L.: The flow of human crowds. Annual review of fluid mechanics **35**(1) (2003) 169–182
15. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Physical review E **51**(5) (1995) 4282
16. Boyle, P., Frean, M.: Dependent gaussian processes. Advances in neural information processing systems **17** (2005) 217–224
17. Tay, M.K.C., Laugier, C.: Modelling smooth paths using gaussian processes. In: Field and Service Robotics, Springer (2008) 381–390
18. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: Computer Vision, 2009 IEEE 12th International Conference on, IEEE (2009) 261–268

19. Treuille, A., Cooper, S., Popović, Z.: Continuum crowds. In: ACM Transactions on Graphics (TOG). Volume 25., ACM (2006) 1160–1168
20. Antonini, G., Venegas, S., Thiran, J.P., Bierlaire, M.: A discrete choice pedestrian behavior model for pedestrian detection in visual tracking systems. In: Advanced Concepts for Intelligent Vision Systems, ACIVS 2004. Number EPFL-CONF-87109, IEEE (2004)
21. Antonini, G., Bierlaire, M., Weber, M.: Discrete choice models of pedestrian walking behavior. Transportation Research Part B: Methodological **40**(8) (2006) 667–687
22. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. Pattern Analysis and Machine Intelligence, IEEE Transactions on **30**(2) (2008) 283–298
23. Ziebart, B.D., Ratliff, N., Gallagher, G., Mertz, C., Peterson, K., Bagnell, J.A., Hebert, M., Dey, A.K., Srinivasa, S.: Planning-based prediction for pedestrians. In: Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on, IEEE (2009) 3931–3936
24. Henry, P., Vollmer, C., Ferris, B., Fox, D.: Learning to navigate through crowded environments. In: Robotics and Automation (ICRA), 2010 IEEE International Conference on, IEEE (2010) 981–986
25. Ziebart, B.D., Maas, A.L., Bagnell, J.A., Dey, A.K.: Maximum entropy inverse reinforcement learning. In: AAAI. (2008) 1433–1438
26. Levine, S., Popovic, Z., Koltun, V.: Nonlinear inverse reinforcement learning with gaussian processes. In: Advances in Neural Information Processing Systems. (2011) 19–27
27. Thompson, S., Horiuchi, T., Kagami, S.: A probabilistic model of human motion and navigation intent for mobile robot path planning. In: Autonomous Robots and Agents, 2009. ICARA 2009. 4th International Conference on, IEEE (2009) 663–668
28. Luber, M., Stork, J.A., Tipaldi, G.D., Arras, K.O.: People tracking with human motion predictions from social forces. In: Robotics and Automation (ICRA), 2010 IEEE International Conference on, IEEE (2010) 464–469
29. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 935–942
30. Leal-Taixé, L., Pons-Moll, G., Rosenhahn, B.: Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, IEEE (2011) 120–127
31. Alahi, A., Ramanathan, V., Fei-Fei, L.: Socially-aware large-scale crowd forecasting. In: CVPR. (2014)
32. Kretzschmar, H., Kuderer, M., Burgard, W.: Learning to predict trajectories of cooperatively navigating agents. In: Robotics and Automation (ICRA), 2014 IEEE International Conference on, IEEE (2014) 4015–4020
33. Yi, S., Li, H., Wang, X.: Understanding pedestrian behaviors from stationary crowd groups. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3488–3496
34. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. IEEE Transactionson Pattern Analysis and Machine Intelligence (2) (2008) 267–282
35. Pirsiavash, H., Ramanan, D., Fowlkes, C.C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: CVPR. (2011)

36. Alahi, A., Boursier, Y., Jacques, L., Vandergheynst, P.: A sparsity constrained inverse problem to locate people in a network of cameras. In: 2009 16th International Conference on Digital Signal Processing, IEEE (2009) 1–7
37. Alahi, A., Jacques, L., Boursier, Y., Vandergheynst, P.: Sparsity driven people localization with a heterogeneous network of cameras. Journal of Mathematical Imaging and Vision (2011) 1–20
38. Roshan Zamir, A., Dehghan, A., Shah, M.: GMCP-Tracker: Global multi-object tracking using generalized minimum clique graphs. In: Proceedings of the European Conference on Computer Vision (ECCV). (2012)
39. Xiang, Y., Alahi, A., Savarese, S.: Learning to track: Online multi-object tracking by decision making. In: International Conference on Computer Vision (ICCV). (2015) 4705–4713
40. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: MOTChallenge 2015: Towards a benchmark for multi-target tracking. arXiv:1504.01942 [cs] (April 2015) arXiv: 1504.01942.
41. Amer, M.R., Xie, D., Zhao, M., Todorovic, S., Zhu, S.C.: Cost-sensitive top-down / bottom-up inference for multiscale activity recognition. In: ECCV. (2012)
42. Shu, T., Xie, D., Rothrock, B., Todorovic, S., Chun Zhu, S.: Joint inference of groups, events and human roles in aerial videos. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2015)
43. Alahi, A., Bierlaire, M., Vandergheynst, P.: Robust real-time pedestrians detection in urban environments with low-resolution cameras. Transportation research part C: emerging technologies **39** (2014) 113–128
44. Alahi, A., Bierlaire, M., Kunt, M.: Object detection and matching with mobile cameras collaborating with fixed cameras. In: Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 2008. (2008)
45. Trautman, P., Krause, A.: Unfreezing the robot: Navigation in dense, interacting crowds. In: Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on, IEEE (2010) 797–803