Point-based Path Prediction from Polar Histograms

Pasquale Coscia, Francesco Castaldo
and Francesco A.N. Palmieri
Dipartimento di Ingegneria Industriale e dell'Informazione
Seconda Università degli Studi di Napoli
Aversa (CE), Italy
{pasquale.coscia, francesco.castaldo,
francesco.palmieri}@unina2.it

Lamberto Ballan, Alexandre Alahi
and Silvio Savarese
Department of Computer Science
Stanford University
California, USA
{lballan, alahi, ssilvio}@cs.stanford.edu

***Abstract*—We address the problem of modeling complex target behavior using a stochastic model that integrates object dynamics, statistics gathered from the environment and semantic knowledge about the scene. The method exploits prior knowledge to build point-wise polar histograms that provide the ability to forecast target motion to the most likely paths. Physical constraints are included in the model through a ray-launching procedure, while semantic scene segmentation is used to provide a coarser representation of the most likely crossable areas. The model is enhanced with statistics extracted from previously observed trajectories and with nearly-constant velocity dynamics. Information regarding the target's destination may also be included steering the prediction to a predetermined area. Our experimental results, validated in comparison to actual targets' trajectories, demonstrate that our approach can be effective in forecasting objects' behavior in structured scenes.**

I. Introduction

Technological advances and increasingly accurate prediction models have allowed to reach high levels of autonomy for all those agents, such as self-driving cars or robots, which need to operate in complex environments where there could be unpredictable behaviors. Short-term predictions are not suitable in those contexts where there are many targets, because they typically do not take into account the final destination and the possible intentions of other agents. Most analyses tend to be rather unrealistic.

Nevertheless, the targets monitored within urban scenarios typically follow preferred directions in specific areas providing precious information that could be exploited, for example, by visual surveillance systems. Although the prior knowledge on the monitored scene can be incomplete or inaccurate, it may provide useful information about the possible actions that a moving agent may undertake. Such a knowledge can surely favor the design of long-term path prediction systems that go beyond predictions of only near future.

The path prediction task has been studied for many years and has a wealth of applications, such as:

- *social robotics*: a moving robot needs to estimate the future actions of the targets around it in order to safely operate in the same area and to reduce the interference with human activities.
- *visual surveillance and event recognition*: ideally we could predict dangerous activities before they happen.

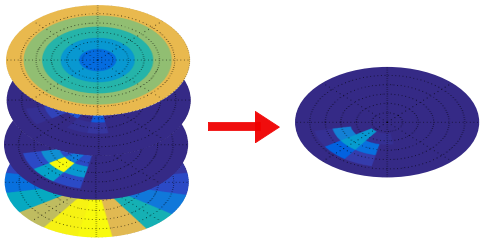


Fig. 1: The stack of polar histograms for the proposed framework. The predicted velocity value $\boldsymbol{v}_k$ is a sample from the resulting histogram on the right.

- *virtual environments*: simulate realistic paths for urban planning and manage emergency situations.
- *public spaces mobility*: learn how common public areas (*e.g.*, airport terminals, shopping malls, *etc.*) are explored by high density crowds, and use this knowledge to better design those spaces (*e.g.*, by modifying the number of floors, the positions of the exits, *etc.*).

In most practical scenarios all we have to work with is noisy, or partial, prior information. For this reason, the uncertainty related to the targets, or to the semantic elements, makes quite often necessary to use probabilistic models to forecast realistic paths. Moreover, goals or mid-level goals, are the key element of the majority of the works on this topic. We show how the missing information about the target's destination, exploiting a limited prior knowledge, may enable to get comparable results as in the presence of the goal.

Our work aims at predicting the future actions of agents moving in a crowded scene. Here our focus is to model the interactions between the target and the scene (*i.e.*, human-scene interactions).The main contributions of this paper are: 1) a fine-level model to forecast realistic paths for different types of targets, 2) the integration of semantic elements and a probabilistic dynamic model. The proposed model shows robust features that make it also suitable for a real-time coarse path prediction.

The paper is organized as follows. In Section 2, we summarize the main recent works on this topic. Section 3 describes in details the elements of our framework. In Section 4, we

investigate the accuracy of path prediction with experimental results, both qualitative and quantitative. Finally, we present conclusions in Section 5.

II. Related Work

Our model leverages trajectory data gathered from the analyzed scene and therefore it falls within the broad area of trajectory-based activity analysis models [1], [2], [3].

Many other lines of research aim at resolving a similar task from different perspectives. For example, in activity recognition [4], [5], [2] the objective is to localize and label agents' actions in the observed scene. Those activities may be very different and may span from individual actions, such as pedestrians walking in the scene [4], or speaking on the phone [5], to complex activities involving groups of people [6], [7].

Among all the approach for temporal modeling of activities, various filtering methods have been proposed, such as Kalman and particle filters [8], [9], [10]. In [11], tracking is also used to identify objects of interest for activity recognition, and in [4] the two tasks are jointly solved, resulting in a more robust framework. In motion planning [8], [12], models are built to guide the path of an agent by leveraging the prior knowledge about its final goal.

Another interesting line of research involves the discovery of human-human interactions in crowded environments [13], [14], [15], which leads to better activity recognition models. These approaches often account for the human behaviors by means of "social" forces [16] which may measure internal motivations of the individuals to perform an action (*e.g.*, attraction from possible goals, repulsion from walls, *etc.*). In [17], a large dataset of trajectories collected in a train station is used to learn how to predict the destinations of moving pedestrians. Other works show also that prior knowledge of goals yields better human-activity recognition and tracking [18], [12]. In [19], [20], behaviors and pairwise interactions are labeled and used to evaluate abnormal or dangerous situations happening in crowded scenes as busy streets or ports.

Some recent work [21], [22], [23], [24] has focused on predicting unobserved future actions. Activity prediction (or forecasting) may not rely on complete observations of the targets as happens for activity recognition. In [21] the forecasting of trajectory-based human activities is achieved by using inverse optimal control and semantic scene labeling. Their approach requires a prior knowledge about the final goal of the target, and their models are not sensitive to the target class. An extension of this work can be found in [23], where a large collection of videos is used to build a model which predicts the most likely future of generic agents (*e.g.*, a car) in the scene. This approach also yields a visual "hallucination" of future likely events on top of the scene. The major drawback or their approach is that they strongly focus on predicting the future appearance and shape of the target and their results are mostly related to a single car-road scenario.

Another interesting approach to the task can be found in [24], where objects of the scene are regarded as "dark matter", emanating an energy that can both attract (*e.g.*, for vending machines) or repulse (*e.g.*, for grass or buildings) the humans during their activities.

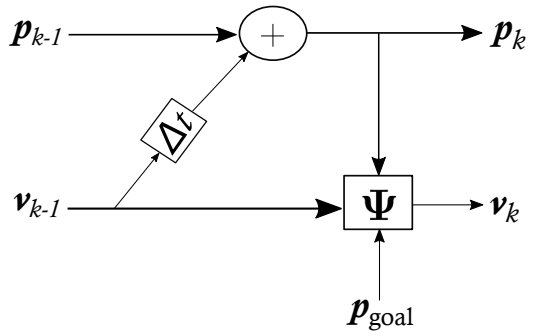


Fig. 2: Graphical representation of the dynamic model.

III. The Model

The state vector at discrete time k of a moving target is $\boldsymbol{X}_k = (\boldsymbol{p}_k, \boldsymbol{v}_k)$ where $\boldsymbol{p}_k = (p_{x_k}, p_{y_k})$ and $\boldsymbol{v}_k = (v_{x_k}, v_{y_k})$ are position and velocity in 2D. The dynamic model is

$$\begin{cases} \boldsymbol{p}_k = \boldsymbol{p}_{k-1} + \boldsymbol{v}_{k-1}\Delta t, \\ \boldsymbol{v}_k \sim \boldsymbol{\Psi}(\boldsymbol{v}_k|\boldsymbol{v}_{k-1}, \boldsymbol{p}_k, \boldsymbol{p}_{goal}), \end{cases} \tag{1}$$

where the conditional density $\boldsymbol{\Psi}(.)$ depends on four independent factors as

$$\begin{aligned} &\boldsymbol{\Psi}(\boldsymbol{v}_k|\boldsymbol{v}_{k-1}, \boldsymbol{p}_k, \boldsymbol{p}_{goal}) \propto \\ &\quad \boldsymbol{S}(\boldsymbol{v}_k|\boldsymbol{p}_k) \cdot \boldsymbol{O}(\boldsymbol{v}_k|\boldsymbol{p}_k) \cdot \boldsymbol{V}(\boldsymbol{v}_k|\boldsymbol{v}_{k-1}) \cdot \boldsymbol{G}(\boldsymbol{v}_k|\boldsymbol{p}_k, \boldsymbol{p}_{goal}), \end{aligned} \tag{2}$$

where $\boldsymbol{S}(\boldsymbol{v}_k|\boldsymbol{p}_k)$ is the *Semantic pdf* that accounts for the structural knowledge about the scene, $\boldsymbol{O}(\boldsymbol{v}_k|\boldsymbol{p}_k)$ is the *Observation pdf* that carries statistical knowledge from previously observed trajectories, $\boldsymbol{V}(\boldsymbol{v}_k|\boldsymbol{v}_{k-1})$ is the *Velocity pdf* that models the intrinsic evolution of the objects' velocity and $\boldsymbol{G}(\boldsymbol{v}_k|\boldsymbol{p}_k, \boldsymbol{p}_{goal})$ is the *Goal pdf* that steers the velocity towards a possible desired destination. The model is depicted in Fig. 2 and each factor will be described in more detail in the following. The velocity generative model Ψ and its factors are discrete versions of their continuous counterparts and are described by polar histograms (PHs). More specifically, at each pixel position $\boldsymbol{p}$, known previous velocity $\boldsymbol{v}_{k-1}$ and goal position $\boldsymbol{p}_{goal}$,

$$\begin{aligned} &\Psi(\rho_i, \theta_j|\boldsymbol{v}_{k-1}, \boldsymbol{p}, \boldsymbol{p}_{goal}) \propto \\ &\quad S(\rho_i, \theta_j|\boldsymbol{p}) O(\rho_i, \theta_j|\boldsymbol{p}) V(\rho_i, \theta_j|\boldsymbol{v}_{k-1}) G(\rho_i, \theta_j|\boldsymbol{p}, \boldsymbol{p}_{goal}), \end{aligned} \tag{3}$$

$(\rho_i, \theta_j) = (\Delta\rho(i-1), \frac{2\pi}{M}(j-1)), i = 1...N+1; j = 1...M$. We fix $\Delta\rho$ to ρ_{max}/N. Our model takes also into account the possibility of the target to stop moving. The velocity factorized model is depicted as the set of the four polar histograms in Fig. 1 with $(N, M) = (5, 16)$, and in Fig. 4 for $(N, M) = (4, 8)$.

A. Semantic PH

The purpose of the Semantic factor is twofold: (a) to avoid that the target impacts on the obstacles; (b) to measure the probability that certain areas are more likely to be crossed then

others. For example, in a top-down view, pedestrians, bikers, cars cannot certainly go through buildings or obstructions. At the same time they are more likely to cross sidewalks, bike paths and streets, respectively. Therefore, given an object class, after a semantic analysis of the scene, we can predetermine how likely it is that certain areas will be crossed.

The first step in the obtaining a Semantic PH is to assign to each pixel $\boldsymbol{p}$ a class label. The alphabet could be $\mathcal{C} = \{sidewalk, road, roundabout, grass, tree, building, obstacle\}$ with a "desirability" value assigned to each element $\mathcal{D} = \{d_{sid}, d_{roa}, d_{rou}, d_{gra}, d_{tre}, d_{bui}, d_{obs}\}$ with $0 \leq d_{\cdot} \leq 1$. For example $d_{obs} = d_{tre} = 0$, because no trajectory can cross an obstacle or a tree; for a biker $d_{roa} > d_{sid} > d_{gra}$, meaning that a biker is more likely to ride on a road than on a sidewalk, or on grass. The values on set $\mathcal{D}$ can be estimated by segmenting images and by counting the number of trajectories that cross that type of region in a given context and for a given object class. In the following, for a biker of the first scenario (see Section IV), we have estimated $\mathcal{D} = \{0.0576, 0.9391, 0, 0.0034, 0, 0, 0\}$. Figure 5c shows an example of a desirability map $\boldsymbol{D}(\boldsymbol{p})$, where targets labeled as *bikers*, prefer to move on streets more than on sidewalks.

The semantic PH is obtained by quantizing $\boldsymbol{S}(\boldsymbol{v}|\boldsymbol{p})$, which is how the next velocity vector is likely to behave at a given pixel position $\boldsymbol{p}$ as a consequence of the semantic map that surrounds it. The function is estimated by means of a sort of *ray-launching* procedure as shown in Fig. 3. Assuming that a maximum speed is set to $|\boldsymbol{v}|_{max}$, the maximum reachable radius is $\rho_{max} = |\boldsymbol{v}|_{max}\Delta t$. If a beam is launched in each direction θ, *i.e.*, a target moves along that ray, we want to measure cumulatively the difficulty to cross the different areas by using our desirability map. We define first a *resistivity* map $\boldsymbol{R}(\boldsymbol{p}) = 1 - \boldsymbol{D}(\boldsymbol{p})$ and estimate at each $\boldsymbol{p}$ the integral on ray paths that starts from $\boldsymbol{p}$ and travels radially up to ρ

$$z(\rho, \theta; \boldsymbol{p}) = \min(1, \int_0^{\rho} \boldsymbol{R}(r, \theta; \boldsymbol{p}) dr), \quad 0 < \rho < \rho_{max}. \quad (4)$$

Here $\boldsymbol{R}(r, \theta; \boldsymbol{p})$ is the resistivity map expressed in polar coordinated with origin in $\boldsymbol{p}$. The notation $\min(1, .)$ expresses the saturation effect (to one) that we have when we hit obstacles, or cross very undesirable areas: at locations where $z(\rho, \theta; \boldsymbol{p}) = 1$ the ray cannot arrive; similarly the paths to the locations where $z(\rho, \theta; \boldsymbol{p}) \simeq 0$ are relatively free. This is translated in a semantic polar distribution as

$$\boldsymbol{S}(\rho, \theta|\boldsymbol{p}) \propto 1 - z(\rho, \theta; \boldsymbol{p}). \quad (5)$$

The Semantic histogram is obtained by using a finite number of directions θ, quantizing ρ and normalizing the result.

B. Observation PH

The Observation PH contains the velocity statistics extracted from the trajectories in the training set. At each pixel location $\boldsymbol{p}$ we count the number of times a velocity vector happens for each (ρ, θ). To better condition the statistics we use a weighted sum of the statistics from the neighboring pixels:

$$\boldsymbol{O}(\boldsymbol{v}|\boldsymbol{p}) = \sum_{i=0}^{N} w_{D_i} \boldsymbol{O}(\boldsymbol{v}|\boldsymbol{p}_i), \quad w_{D_i} = (1-r)^{D_i} \quad (6)$$

where N is the number of adjacent pixels, r is fixed to 0.8 and D_i is distance from the adjacent pixel.

In the regions where no trajectories are present, we assume a uniform distribution.

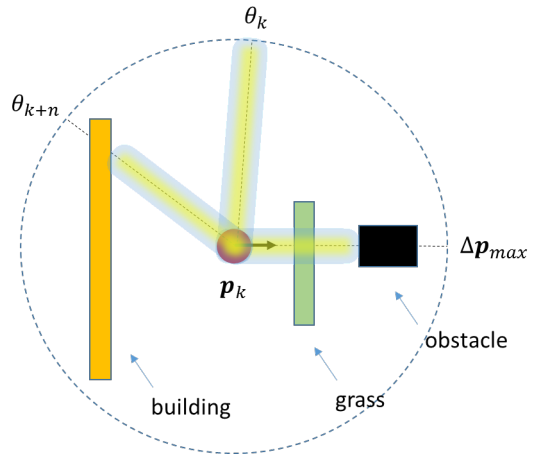


Fig. 3: Illustration of the *ray-launching* procedure. We imagine a laser shooting in several directions. The ray stops when obstacles (non-crossable semantic classes), or the maximum possible displacement from the initial position, are reached.

C. Velocity PH

In this work, a nearly-constant velocity model is used. This is justified by the fact that it is highly unlikely that a target changes suddenly its own direction and it is generally inclined to keep its previous velocity. In other words, the velocity at the step k can be viewed as a noisy version of the velocity at the step $k-1$. Making a classical Gaussian assumption we have

$$\boldsymbol{V}(\boldsymbol{v}_k|\boldsymbol{v}_{k-1}) \sim \mathcal{N}(\boldsymbol{v}_k; \boldsymbol{v}_{k-1}, \sigma^2 I_2). \quad (7)$$

The variance σ is one of the free parameters for our framework. The histogram for the velocity vector in polar coordinates, for accurate quantization, is computed (numerically) evaluating the integral

$$V(\rho_i, \theta_j|v_{x_{k-1}}, v_{y_{k-1}}) = \iint_{\Omega_{i,j}} \frac{\rho}{2\pi\sigma^2} e^{-\frac{(\rho cos\theta - v_{x_{k-1}})^2 + (\rho sin\theta - v_{y_{k-1}})^2}{2\sigma^2}} d\rho \, d\theta \quad (8)$$

where $\Omega_{i,j} = \{(\rho, \theta) : \rho_i \leq \rho \leq \rho_{i+1}, \theta_j \leq \theta \leq \theta_{j+1}\}, i = 1, ..., N+1; j = 1, ..., M$.

D. Goal PH

The goal, if known, is taken into account with a further distribution, $\boldsymbol{G}(\boldsymbol{v}_k|\boldsymbol{p}_k, \boldsymbol{p}_{goal})$, where $\boldsymbol{p}_{goal}$ is the goal position. We use a von Mises distribution with mean $\theta_g = \angle(\boldsymbol{p}_k, \boldsymbol{p}_{goal})$, *i.e.*, the angle between the two vectors representing the target

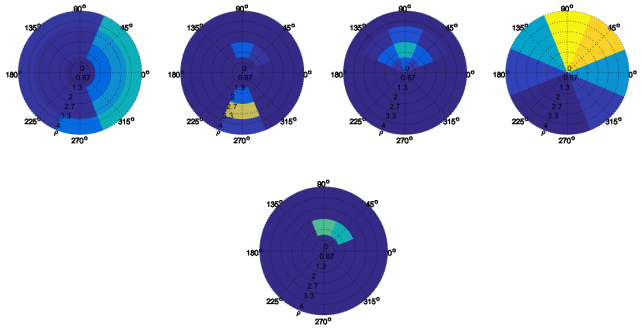


Fig. 4: The first line shows an example of the four polar histograms of our framework which represent the semantic, the observations, the gaussian velocity and the goal direction. The second line shows the resulting polar histogram. The v_{max} value is extracted as the maximum speed of the trajectories of the training set.

and the goal positions. This distribution gives a nice approximation of a wrapped normal distribution around the circle. The distribution does not depend on ρ and its quantized histogram is obtained by computing (numerically) the integral

$$G(\theta_i|\theta_g,\kappa) = \frac{1}{Z}\int_{\theta_i-\Delta\theta}^{\theta_i+\Delta\theta} \frac{e^{\kappa cos(\theta-\theta_g)}}{2\pi I_0(\kappa)} d\theta, \quad i = 1,...,M. \quad (9)$$

where Z is a normalization factor, $I_0(\kappa)$ is the modified Bessel function of zero order and κ is the concentration parameter.

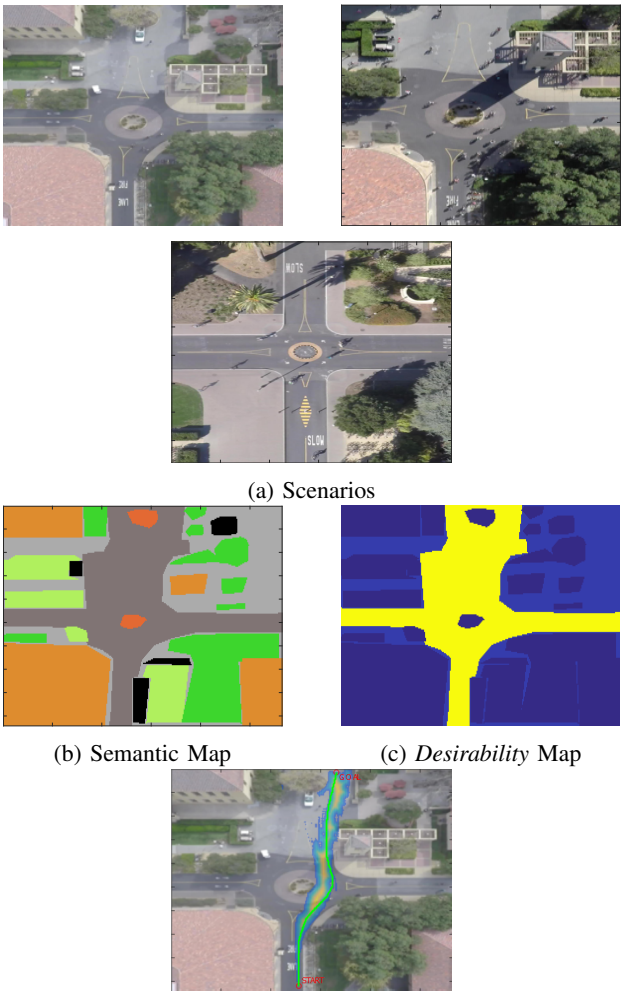
E. Fusing the histograms and selecting the paths

The resulting polar histogram is calculated as the normalized product of the four histograms described above and depicted in Fig. 1 and Fig. 4. Different paths are generated by sampling from the final histogram.

To remove spurious occurrences, our final set of trajectories is obtained from a further selection phase. Using previously measured trajectories, we compute also a *Popularity map*. For each pixel of a known scene, we compute from a training set, the number of times trajectories cross it. For each generated trajectory then we compute the cumulated popularity score by summing all popularity values of the crossed pixels. Only the most popular trajectories are retained.

IV. Experiments

For our experiments we use a new dataset of urban scenes introduced in [25]. The dataset provides a rich set of trajectories of different targets' classes (pedestrians, bikers, *etc.*), along with semantic annotations of the scenes (the following 10 classes are used: *road, roundabout, sidewalk, grass, tree, bench, building, bike rack, parking lot, background*). As a proof-of-concept, we focus our attention on three complex urban scenarios drawn by the dataset, depicted in Fig. 5a. In this work we reduce some semantic classes to a generic



(a) Scenarios

(b) Semantic Map

(c) *Desirability* Map

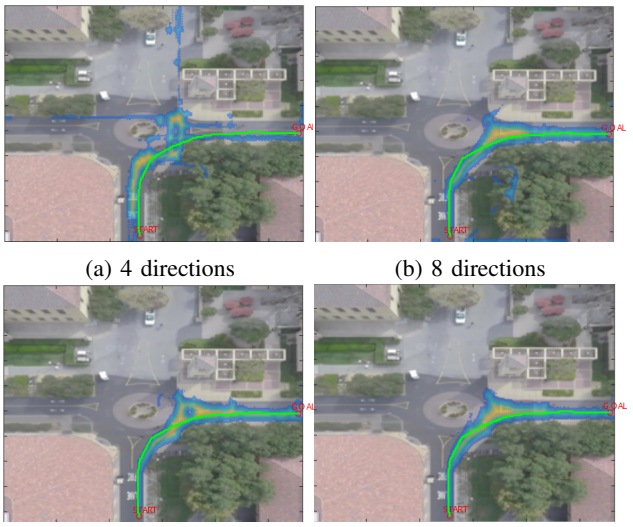
(d) Heat Map

Fig. 5: The figure shows (a) the three selected scenarios of the dataset, and for the first scenario (b) the corresponding semantic map, (c) the desirability map for the target class *biker*, (d) the heat map, with the destination information, for a ground-truth (in green). The considered semantic classes are: sidewalk, road, roundabout, grass, tree, building and obstacle.The *desirability map* represents the base for the *ray-launching* procedure that takes into account the number of trajectories that cross each semantic class.

semantic class, labeled as *obstacle*, since they do not impact on the chosen scenarios.

Our approach requires to compare trajectories of different length, therefore we use the modified Hausdorff distance (MHD) [26] as a metric of distance between the generated trajectories and the ground truth. For our experiments, the process of path generation is stopped when the target reaches a fixed-size area around the goal of 3x3 pixels, or when it reaches one of the (manually-annotated) exits. Moreover, in this work we focus on *bikers* trajectories, but ideally we could analyze any kind of moving targets, such as pedestrians, cars, skaters, *etc.*

To initialize the prediction process, we pick both starting and



(a) 4 directions (b) 8 directions

(c) 12 directions (d) 16 directions

Fig. 6: Qualitative experiments showing the heat maps of the predicted trajectories at different resolutions (*i.e.*, the number of directions) for a selected ground-truth. We have considered all the available elements, including the direction of the goal. The parameters σ and κ have been set to 0.45 and 2, respectively. In green, the ground-truth trajectory is shown.

Resolution	Mean Error (MHD)
4	21.4567
8	9.2942
12	6.2889
16	4.1391

TABLE I: This table shows the mean errors (MHD) of the most popular trajectories depicted in Fig.6 with respect to the selected ground-truth.

final points (if the goal is known) from a randomly-selected trajectory of the test set. At the first step, the $\boldsymbol{V}$ distribution is initialized with a uniform distribution since we suppose to know the initial position, but not the target's velocity. In this manner, the velocity acquired by the target, at the next step, will depend only on the semantic and statistics data, and on the goal, if known.

A. Qualitative experiments

Fig 6 shows the predicted paths varying the number of directions, assuming to know the goal. Four directions are not enough for a rich description of movements. Increasing the number of directions, the generated trajectories are close enough to the ground-truth path. Those results are confirmed in the Table I. We maintain fixed the number of directions to 16 for all the experiments.

B. Quantitative experiments

When the goal is not known, we compare our approach with two baselines: a random walk and a constant velocity model. We have used the 80% of the dataset as training set and the rest as test set. In order to have a fair comparison,

	N = # of trajectories				
Mean Error (MHD)	10	20	30	40	50
Random Walk	55.1976	53.7375	58.6903	33.1718	37.7118
CV Model	48.6914	31.3831	44.5100	63.1029	45.1445
Ours	29.5719	25.9621	25.1530	24.8934	24.6881

TABLE II: Mean Error (MHD) to evaluate the accuracy parameter. The value has been also reported for two baselines: a random walk and a constant velocity model.

we store 100 trajectories generated by both baselines and our approach, and we evaluate the accuracy in terms of average error with the MHD metric. In particular, we compare a subset of all the ground truth paths originating within a 5x5 pixels region around the starting point, and the N most popular paths (the popularity is evaluated in the way described in Sec. III-E) drawn from both baselines and our framework. For each ground truth path we calculate the mean error for all the paths, and then we average all those errors to obtain a single accuracy number. As shown in Tab. II, we easily verify that our framework outperforms the baselines.

C. Precision and accuracy evaluation

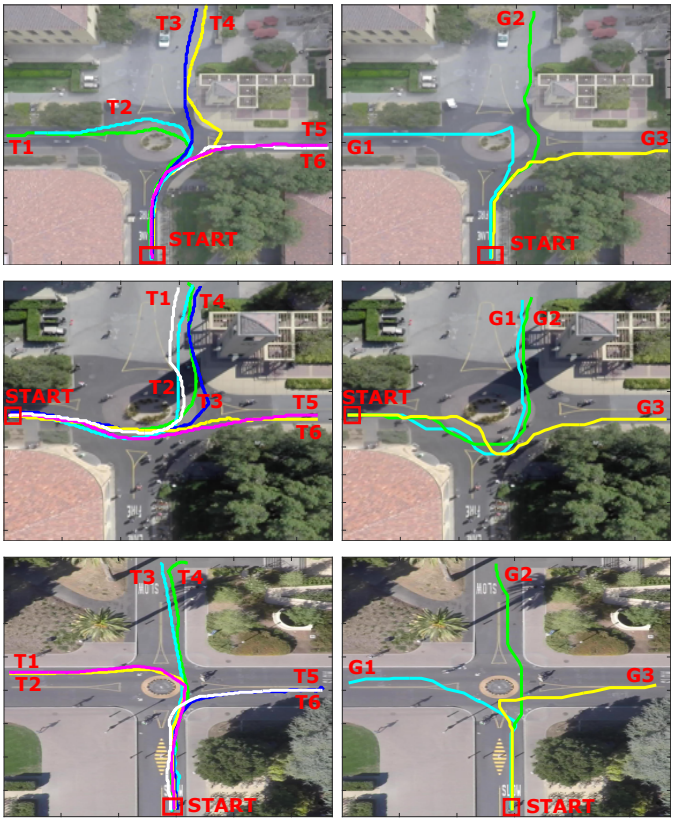
We also propose another evaluation, where 6 representative trajectories (depicted in Fig. 7a) have been picked from the ground truth and compared against all the predicted paths. Each generated path is associated with the closest of the 6 trajectories, and we build an histogram by counting the associations. The numerical results, for the first scenario, shown in Table III, confirm that our framework uniformly distributes the generated paths on all the chosen trajectories, except for trajectory *T4* that is never matched due to its atypical behavior.

Table IV shows the mean errors (MHD) evaluated with respect to a selected ground-truth when one or more elements are removed from the framework. To calculate the mean error, in this case we considered all the 100 trajectories generated by our framework. As expected, the predicted trajectories are mainly influenced by the final destination. However, the observed trajectories provide an effective prior knowledge to predict the real trajectory of the selected target when the goal is not known.

Finally, in Fig. 8, we report the error we get by selecting different values of σ and κ for a number of ground-truths of the three scenarios. We assume here to know the goal. As shown, lower values of σ do not allow the target to reach its final destination while intermediate values of the concentration parameter κ (2 and 4) reduce the error curves compared to the other values.

V. Conclusion

In this paper we have focused on complex human-scene interactions that are learned from trajectory data of the scene and leveraged to forecast plausible paths of a target in urban



(a) (b)

Fig. 7: The figure shows (a) the trajectories drawn from the scenarios with the same starting area, (b) some of the most popular generated paths.

Observed trajectory	N = # of trajectories 10	20	30	40	50
T1	20%	15%	16.66%	15%	12%
T2	20%	20%	23.34%	22.5%	24%
T3	30%	25%	20%	20%	20%
T4	0%	0%	0%	0%	0%
T5	20%	15%	13.34%	20%	18%
T6	10%	25%	26.66%	22.5%	26%

TABLE III: Percentage of the generated trajectories which are closest to the observed trajectories shown in Fig. 7a related to the first scenario.

Elements	Mean Error (MHD)	Elements	Mean Error (MHD)
All	2.0456	All-$\{\boldsymbol{O},\boldsymbol{V}\}$	3.5639
All-$\boldsymbol{O}$	13.5971	All-$\{\boldsymbol{O},\boldsymbol{G}\}$	29.3877
All-$\boldsymbol{V}$	3.4109	All-$\{\boldsymbol{V},\boldsymbol{G}\}$	24.6377
All-$\boldsymbol{G}$	9.5532		

TABLE IV: Mean errors (MHD) for a selected ground-truth of the first scenario when some elements are removed. The parameters σ and κ have been set to 0.45 and 2, respectively.

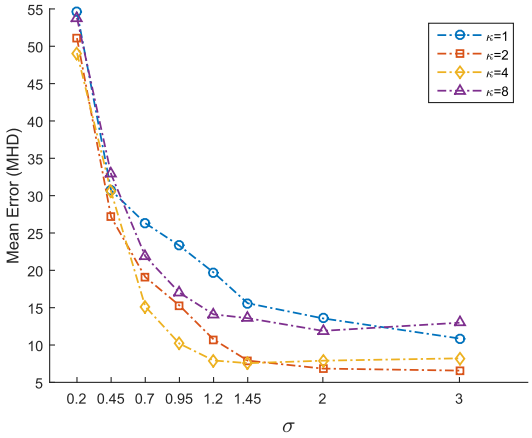


Fig. 8: This figure shows the mean errors (MHD) with different values of σ and κ with respect to a randomly selected ground-truth. For higher values of σ, the $\boldsymbol{V}$ distribution does not affect the error since it tends to be a uniform distribution.

scenarios. The next steps are to focus on a *macro*-pixel approach rather than a *per*-pixel approach, and mainly on the interactions between targets (*i.e.*, human-human interactions). In our opinion the joint estimation of these two types of interactions is the key to build a system that is able to forecast realistic paths, given only the starting point and the scene statistics. Another interesting line of research involves the transfer of the information gathered from a scene to other "similar" scenes in terms of semantics: the idea here is to investigate how general behaviors (as for instance the way humans move on sidewalks or near roundabouts) can be transfered to new unseen scenes (*e.g.*, scenes for which we do not have any trajectory data at our disposal).

Acknowledgment

This project has been partially sponsored by a grant from the italian MIUR (Ministero dell'Istruzione e della Ricerca Scientifica) through CNIT (Consorzio Interuniversitario per le Telecomunicazioni), PON03PE-00185-1,2 (Prog. MAR.TE.).

References

[1] B. Morris and M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, pp. 1114–1127, 2008.

[2] X. Wang, K. T. Ma, G.-W. Ng, and E. Grimson, "Trajectory analysis and semantic region modeling using a nonparametric bayesian model," in *CVPR*, 2008.

[3] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *CVPR*, 2012.

[4] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *ECCV*, 2012.

[5] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu, "Cost-sensitive top-down / bottom-up inference for multiscale activity recognition," in *ECCV*, 2012.

[6] T. Lan, W. Yang, Y. Wang, and G. Mori, "Beyond actions: Discriminative models for contextual group activities," in *NIPS*, 2010.

[7] W. Choi, Y. W. Chao, C. Pantofaru, and S. Savarese, "Discovering groups of people in images," in *ECCV*, 2014.

[8] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. MIT Press, 2005.

[9] Y. Bar-Shalom, T. Kirubarajan, and X.-R. Li, *Estimation with Applications to Tracking and Navigation: Theory Algorithms and Software*. John Wiley & Sons, 2002.

[10] Y. Bar-Shalom, P. Willett, and X. Tian, *Tracking and Data Fusion: A Handbook of Algorithms*. YBS Publishing, 2011.

[11] X. Wang, X. Ma, and E. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 3, pp. 539–555, March 2009.

[12] H. Gong, J. Sim, M. Likhachev, and J. Shi, "Multi-hypothesis motion planning for visual object tracking," in *Proceedings of the 2011 International Conference on Computer Vision*, ser. ICCV '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 619–626. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2011.6126296

[13] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *ICCV*, 2009.

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

[15] L. Leal-Taixe, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese, "Learning an image-based motion context for multiple people tracking," in *CVPR*, 2014.

[16] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *CVPR*, 2009.

[17] A. Alahi, V. Ramanathan, and L. Fei-Fei, "Socially-aware large-scale crowd forecasting," in *CVPR*, 2014.

[18] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *ECCV*, 2008.

[19] F. Castaldo, V. Bastani, L. Marcenaro, F. Palmieri, and C. Regazzoni, "Abnormal vessel behavior detection in port areas based on dynamic bayesian networks," in *17th IEEE International Conference on Information Fusion (FUSION)*, 2014.

[20] F. Castaldo, F. A. N. Palmieri, and C. S. Regazzoni, "Bayesian analysis of behaviors and interactions for situation awareness in transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 313–322, Feb 2016.

[21] K. Kitani, B. Ziebart, J. Bagnell, and M. Hebert, "Activity forecasting," in *ECCV*, 2012.

[22] D. F. Fouhey and C. Lawrence Zitnick, "Predicting object dynamics in scenes," in *CVPR*, 2014.

[23] J. Walker, A. Gupta, and M. Hebert, "Patch to the future: Unsupervised visual prediction," in *CVPR*, 2014.

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

[25] A. Robicquet, A. Alahi, A. Sadeghian, B. Anenberg, J. Doherty, E. Wu, and S. Savarese, "Forecasting social navigation in crowded complex scenes," *CoRR*, vol. abs/1601.00998, 2016.

[26] M. P. Dubuisson and A. K. Jain, "A modified hausdorff distance for object matching," in *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, vol. 1, Oct 1994, pp. 566–568 vol.1.