Inferring “Dark Matter” and “Dark Energy” from Videos

Dan Xie*, Sinisa Todorovic†, and Song-Chun Zhu*

* Center for Vision, Cognition, Learning and Art
Depts. of Statistics and Computer Science
University of California, Los Angeles, USA
xiedan@g.ucla.edu, sczhu@stat.ucla.edu

† School of EECS
Oregon State University, USA
sinisa@onid.orst.edu

Abstract

*This paper presents an approach to localizing functional objects in surveillance videos without domain knowledge about semantic object classes that may appear in the scene. Functional objects do not have discriminative appearance and shape, but they affect behavior of people in the scene. For example, they “attract” people to approach them for satisfying certain needs (*e.g.*, vending machines could quench thirst), or “repel” people to avoid them (*e.g.*, grass lawns). Therefore, functional objects can be viewed as “dark matter”, emanating “dark energy” that affects people’s trajectories in the video. To detect “dark matter” and infer their “dark energy” field, we extend the Lagrangian mechanics. People are treated as particle-agents with latent intents to approach “dark matter” and thus satisfy their needs, where their motions are subject to a composite “dark energy” field of all functional objects in the scene. We make the assumption that people take globally optimal paths toward the intended “dark matter” while avoiding latent obstacles. A Bayesian framework is used to probabilistically model: people’s trajectories and intents, constraint map of the scene, and locations of functional objects. A data-driven Markov Chain Monte Carlo (MCMC) process is used for inference. Our evaluation on videos of public squares and courtyards demonstrates our effectiveness in localizing functional objects and predicting people’s trajectories in unobserved parts of the video footage.*

1. Introduction

This paper considers the problem of localizing functional objects and scene surfaces in surveillance videos of public spaces, such as courtyards and squares. The functionality of objects is defined in terms of force-dynamic effects that they have on human behavior in the scene. For instance, people may move toward certain objects (*e.g.*, food truck, vending machines, and chairs), where they can satisfy their needs (*e.g.*, satiate hunger, quench thirst, or have rest), as illustrated in Fig. 1. Also, while moving, people will tend to avoid non-walkable areas (*e.g.*, grass lawns) and obstacles. In our low-resolution surveillance videos, these functional objects and surfaces cannot be reliably recognized by their appearance and shape. But their presence noticeably affects people’s trajectories. Therefore, by analogy to cosmology, we regard these unrecognizable functional objects as sources of “dark energy”, *i.e.*, “dark matter”, which exert attraction and repulsion forces on people.

Recognizing functional objects is a long standing problem in vision, with slower progress in the past decade, in contrast to impressive advances in appearance-based recognition. One reason is that appearance features generally provide poor cues about the functionality of objects. Moreover, for low-resolution, bird’s-eye-view surveillance videos, considered in this paper, appearance features are not sufficient to support robust object detection. Instead, we analyze human behavior in the video by predicting people’s intents and motion trajectories, and thus localize sources of “dark energy” that drive the scene dynamics.

To approach this problem, we leverage the Lagrangian mechanics (LM) by treating the scene as a physical system. In such a system, people can be viewed as charged particles moving along a mixture of repulsion and attraction energy fields generated by “dark matter”. The classical LM, however, provides a poor model of human behavior, because it wrongly predicts that people always move toward the closest “dark matter”, by the principle of least action.

We extend the classical LM to agent-based LM (ALM), which accounts for human latent intents. Specifically, we make the assumption that people intentionally approach functional objects (to satisfy their needs). This amounts to enabling the charged particles in ALM to become agents who can personalize the strengths of “dark energy” fields by appropriately weighting them. In this way, every agent’s

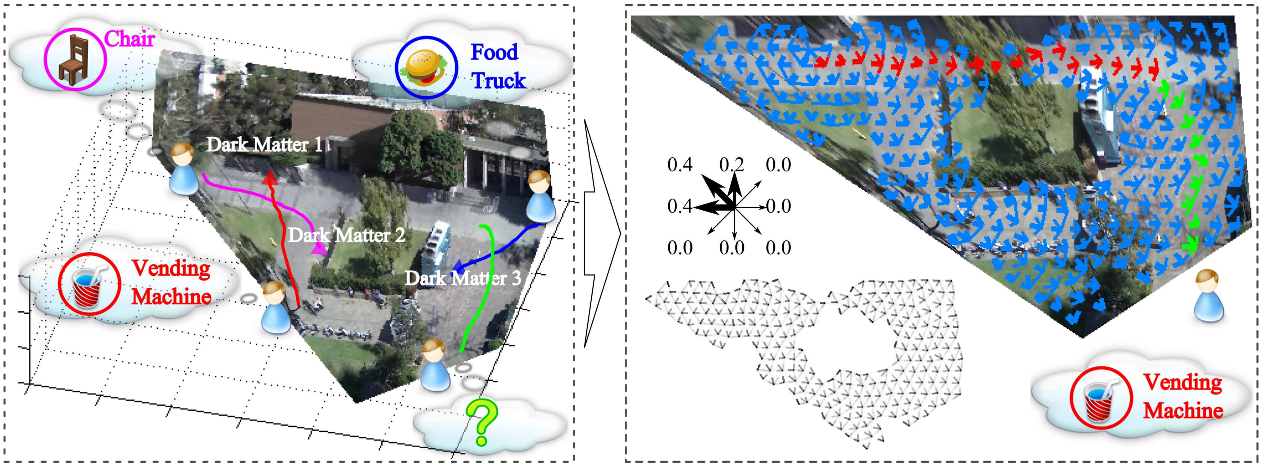

Figure 1. An example video where people driven by latent needs (e.g., hunger, thirst) move toward "dark matter", where these needs can be satisfied (e.g., food truck, vending machine). We analyze human latent intents and trajectories to localize "dark matter". For some people (bottom right person) we observe only an initial part of their trajectory (green). (Right) Our actual results of: (a) inferring a given person's latent intent; (b) predicting the person's full trajectory (red); (c) locating one source of "dark energy" (vending machine); (d) estimating the constraint map of non-walkable areas; and (e) estimating the force field affecting the person (edge thickness indicates magnitude, and below is another visualization of the same force field with "holes" corresponding to our estimates of non-walkable areas).

motion will be strongly driven by the intended "dark matter", subject to "dark energy" fields of the other sources.

Since our focus is on videos of wide public spaces, we expect that people know the layout of obstacles, walkable, and non-walkable areas in the scene, either from previous experience or simply by observing the scene. This allows the agents to globally optimize their trajectories in the attraction energy field of their choice.

Overview: Given a short video excerpt, providing only partial observations of people's trajectories, we predict:

- Locations of functional objects ("dark matter"), S,
- Goals of every person, R,
- People's full trajectories in unobserved video parts, Γ

To facilitate our prediction of S, R, and Γ, we also infer latent constraint map of non-walkable areas, C, and latent "dark energy" fields, $\vec{F}$. Note that providing ground-truth annotations of C and $\vec{F}$ is fundamentally difficult, and thus we do not evaluate the inferred C and $\vec{F}$.

Our first step is feature extraction, which uses the state-of-the-art multitarget tracker of [19] for detecting and tracking people, as well as the low-level 3D scene reconstruction of [26]. While the tracker and 3D scene reconstruction perform well, they may yield noisy results. These noisy observations are used as input features to our model. Uncertainty is handled within the Bayesian framework, which specifies a joint distribution of observable and latent random variables, where observables are input features, and latent variables include locations of "dark matter", people's goals and trajectories, constraint map, and "dark energy" fields. A data-driven Monte Carlo Markov Chain (MCMC) is used for inference [23, 13]. In each iteration, MCMC samples the number and locations of functional objects and people's goals. This, in turn, uniquely identifies "dark energy" fields. Since people are assumed to know the scene layout, every person's full trajectory can be predicted as a globally optimal Dijkstra path on the scene lattice. These predictions are considered in the next MCMC iteration for the probabilistic sampling of the latent variables.

We present experimental evaluation on surveillance videos from the VIRAT [16] and UCLA Courtyard [3] datasets, as well as on our two webcam videos of public squares. The experiments demonstrate high accuracy in locating "dark matter" in various scenes. We also compare our predictions of human trajectories with those of existing approaches. The results show that we improve upon a number of baselines, and outperform the state of the art.

In the sequel, Sec. 2 reviews prior work, Sec. 3 presents our agent-based Lagrangian mechanics, Sec. 4 formulates our model, Sec. 5 specifies our MCMC inference, and Sec. 6 presents our empirical evaluation.

2. Related Work and Our Contributions

Our work is related to three research streams.

Functionality. Recent work focuses on improving object recognition by identifying their functionality. Calculators or cellphones are recognized in [6, 9], and chairs are recognized in [8], based on the close-body context. [24] labels functional scene elements, *e.g.*, parking spaces, by extracting local motion features. We instead predict a person's goal and full trajectory to localize functional objects.

Event prediction and simulation. The work on early prediction of human activities uses dynamic programming

[20], grammars [17], and max-margin classification [10]. For prediction of human trajectories, [11] uses a deterministic vector field of people's movements, while our "dark energy" fields are stochastic. A linear dynamic system of [27, 28] models *smooth* trajectories of pedestrians in crowded scenes, and thus cannot handle sudden turns and detours caused by obstacles, as required in our setting. In graphics, relatively simplistic models of agents are used to simulate people's trajectories in a virtual crowd [14, 15, 18].

Human tracking and planning. The Lagrangian particle dynamics of crowd flows [1, 2] and the optical-flow based dynamics of crowd behaviors [22] do not account for individual human intents. [7] reconstructs an unobserved trajectory part between two observed parts by finding the shortest path. [5] constructs a numeric potential field for robot path planning. Optimal path search of [21], and reinforcement learning and inverse reinforcement learning of [4, 12] explicitly reason about people's goals for predicting human trajectories. However, these approaches critically depend on domain knowledge. For example, [12] estimates a reward of each semantic object class, detected using an appearance-based object detector. These approaches are not suitable for our problem, since instances of the same semantic class (*e.g.*, two grass lawns in Fig. 1) may have different functionality (*e.g.*, people may walk on one grass lawn, but are forbidden to step on the other).

Our contributions:

- Agent-based Lagrangian Mechanics (ALM) for modeling human behavior in an outdoor scene without exploiting high-level domain knowledge.
- We are not aware of prior work on modeling and estimating the force-dynamic functional map of a scene.
- We distinguish human activities in the video by the associated latent human intents, rather than use the common semantic definitions of activity classes.

3. Background: Lagrangian Mechanics

The Lagrangian mechanics (LM) studies particles with mass, m, and velocity, $\dot{\mathbf{x}}(t)$, in time t, at positions $\mathbf{x}(t) = (x(t), y(t))$ in a force field $\vec{F}(\mathbf{x}(t))$ affecting the motion of the particles. The Lagrangian function, $L(\mathbf{x}, \dot{\mathbf{x}}, t)$, summarizes the kinetic and potential energy of the entire physical system, and is defined as $L(\mathbf{x}, \dot{\mathbf{x}}, t) = \frac{1}{2}m\dot{\mathbf{x}}(t)^2 + \int_{\mathbf{x}} \vec{F}(\mathbf{x}(t))d\vec{\mathbf{x}}(t)$. Action is a key attribute of the physical system, and defined as: $\Gamma(\mathbf{x}, t_1, t_2) = \int_{t_1}^{t_2} L(\mathbf{x}, \dot{\mathbf{x}}, t)dt$. The Lagrangian mechanics postulates that the motion of a particle is governed by the Principle of Least Action: $\hat{\Gamma}(\mathbf{x}, t_1, t_2) = \arg\min_{\Gamma} \int_{t_1}^{t_2} L(\mathbf{x}, \dot{\mathbf{x}}, t)dt$.

The classical LM is not directly applicable to our problem, because it considers inanimate objects. We extend LM in two key aspects, deriving the Agent-based Lagrangian mechanics (ALM). In ALM, the physical system consists of a set of force sources. Our first extension enables the particles to become agents with free will to select a particular force source from the set which can drive their motion. Our second extension endows the agents with knowledge about the layout map of the physical system. Consequently, they can globally plan their trajectories so as to efficiently navigate toward the selected force source, by the Principle of Least Action, avoiding known obstacles along the way. These two extensions can be formalized as follows.

Let ith agent choose jth source from the set of sources. Then, ith agent's action, *i.e.*, trajectory is

$$\begin{aligned}&\Gamma_{ij}(\mathbf{x}, t_1, t_2)\\ &= \arg\min_{\Gamma} \int_{t_1}^{t_2} \left[\frac{1}{2}m\dot{\mathbf{x}}(t)^2 + \int_{\mathbf{x}} \vec{F}_{ij}(\mathbf{x}(t))d\vec{\mathbf{x}}(t)\right]dt. \quad (1)\end{aligned}$$

In our setting (people in public areas), it is reasonable to assume that every agent's speed is upper bounded by some maximum speed. Consequently, from (1), we derive:

$$\Gamma_{ij}(\mathbf{x}, t_1, t_2) = \arg\min_{\Gamma} \int_{t_1}^{t_2} ||\vec{F}_{ij}(\mathbf{x}(t))|| \cdot ||\vec{\Delta}\mathbf{x}(t)||dt. \quad (2)$$

Given $\vec{F}_{ij}(\mathbf{x}(t))$, we use the Dijkstra algorithm for finding a globally optimal solution of (2), since the agents can globally plan their trajectories. Note that the end location of the predicted $\Gamma_{ij}(\mathbf{x}, t_1, t_2)$ corresponds to the location of the selected source j. It follows that estimating the agents' intents and trajectories can be readily used for estimating the functional map of the physical system.

4. Problem Formulation

This section specifies our probabilistic formulation of the problem in a "bottom-up" way. We begin with the definitions of observable and latent variables, and then specify their joint probability distribution.

The video shows agents, $A = \{a_i : i = 1, ..., M\}$, and sources of "dark energy", $S = \{\mathbf{s}_j : j = 1, ..., N\}$, occupying locations on a 2D lattice, $\Lambda = \{\mathbf{x} = (x, y) : x, y \in \mathbb{Z}_+\}$. The locations $\mathbf{x} \in \Lambda$ may be walkable or non-walkable, as indicated by a constraint map, $C = \{c(\mathbf{x}) : \forall\mathbf{x} \in \Lambda,\ c(\mathbf{x}) \in \{-1, 1\}\}$, where $c(\mathbf{x}) = -1$, if $\mathbf{x}$ is non-walkable, and $c(\mathbf{x}) = 1$, otherwise. The allowed locations of agents in the scene are $\Lambda_C = \{\mathbf{x} : \mathbf{x} \in \Lambda,\ c(\mathbf{x}){=}1\}$. Below, we define the priors and likelihoods of these variables that are suitable for our setting.

Constraint map. The prior $P(C)$ enforces spatial smoothness using the standard Ising random field: $P(C) \propto \exp[\beta \sum_{\mathbf{x}\in\Lambda, \mathbf{x}'\in\partial\mathbf{x}\cap\Lambda} c(\mathbf{x})c(\mathbf{x}')],\ \beta > 0$.

Dark Matter. The sources of "dark energy", $\mathbf{s}_j \in S$, and characterized by $\mathbf{s}_j = (\boldsymbol{\mu}_j, \Sigma_j)$, where $\boldsymbol{\mu}_j \in \Lambda$ is the location of $\mathbf{s}_j$, and Σ_j is a 2×2 spatial covariance matrix of $\mathbf{s}_j$'s force field. The distribution of S is conditioned on

C, where the total number $N = |S|$ and occurrences of the sources are modeled with the Poisson and Bernoulli pdf's:

$$P(S|C) \propto \frac{\eta^N}{N!} e^{-\eta} \prod_{j=1}^{N} \rho^{\frac{c(\boldsymbol{\mu}_j)+1}{2}} (1-\rho)^{\frac{1-c(\boldsymbol{\mu}_j)}{2}} \tag{3}$$

where parameters $\eta > 0$, $\rho \in (0,1)$, and $c(\boldsymbol{\mu}_j) \in \{-1, 1\}$.

Agent Goals. Each agent $a_i \in A$ can pursue only one goal, i.e., move toward one source $\mathbf{s}_j \in S$, at a time. The agents cannot change their goals until they reach the selected source. If $a_i \in A$ wants to reach $\mathbf{s}_j \in S$, we specify that their relationship $r_{ij} = r(a_i, \mathbf{s}_j) = 1$; otherwise, $r_{ij} = 0$. Note that r_{ij} is piecewise constant over time. The end-moments of these intervals can be identified when a_i arrives at or leaves from $\mathbf{s}_j$. The set of all relationships is $R = \{r_{ij}\}$. The distribution of R is conditioned on S, and modeled using the multinomial distribution with parameters $\boldsymbol{\theta} = [\theta_1, ..., \theta_j, ..., \theta_N]$,

$$P(R|S) = \textstyle\prod_{j=1}^{N} \theta_j^{b_j}, \tag{4}$$

where θ_j is viewed as a prior of selecting $\mathbf{s}_j \in S$, and each $\mathbf{s}_j \in S$ can be selected b_j times to serve as a goal destination, $b_j = \sum_{i=1}^{M} \mathbb{1}(r_{ij} = 1)$, $j = 1, ..., N$.

Repulsion Force. Every non-walkable location $c(\mathbf{x}){=}{-1}$ generates a repulsion Gaussian vector field, with large magnitudes in the vicinity of $\mathbf{x}$, but rapidly falls to zero. The sum of all these Gaussian force fields on Λ forms the joint repulsion force field, $\vec{F}^-(\mathbf{x})$.

Attraction Forces. Each $\mathbf{s}_j \in S$ generates an attraction Gaussian force field, $\vec{F}_j^+(\mathbf{x})$, where the force magnitude, $|\vec{F}_j^+(\mathbf{x})| = G(\mathbf{x}; \boldsymbol{\mu}_j, \Sigma_j)$, is the Gaussian. When $a_i \in A$ selects a particular $\mathbf{s}_j \in S$, a_i is affected by the corresponding cumulative force field:

$$\vec{F}_{ij}(\mathbf{x}) = \vec{F}^-(\mathbf{x}) + \vec{F}_j^+(\mathbf{x}). \tag{5}$$

Note that by the classical LM, all the agents would be affected by a sum of all force fields: $\vec{F}_{\text{classic}}(\mathbf{x}) = \vec{F}^-(\mathbf{x}) + \sum_{n=1}^{N} \vec{F}_j^+(\mathbf{x})$, instead of $\vec{F}_{ij}(\mathbf{x})$.

Note that an instantiation of latent variables C, S, R uniquely defines the force field $\vec{F}_{ij}(\mathbf{x})$, given by (5).

Trajectories. If $r_{ij} = 1$ then a_i moves toward $\mathbf{s}_j$ along trajectory $\Gamma_{ij} = [\mathbf{x}_i, ..., \mathbf{x}_j]$, where $\mathbf{x}_i$ is a_i's starting location, and $\mathbf{x}_j$ is $\mathbf{s}_j$'s location. Γ_{ij} represents a contiguous sequence of locations on Λ_C. The set of all trajectories is $\Gamma = \{\Gamma_{ij}\}$. As explained in Sec. 3, the agents can globally optimize their paths, because they are familiar with the scene map. Thus trajectory, Γ_{ij}, can be estimated from (2) using the Dijkstra algorithm:

$$\Gamma_{ij} = \arg\min_{\Gamma \subset \Lambda_C} \textstyle\sum_{\mathbf{x} \in \Gamma} ||\vec{F}_{ij}(\mathbf{x}(t))|| \cdot ||\overrightarrow{\Delta\mathbf{x}}(t)||. \tag{6}$$

The likelihood $P(\Gamma_{ij}|C, S, r_{ij}{=}1)$ is specified in terms of the total energy that a_i must spend by walking along Γ_{ij} as

$$\begin{array}{c} P(\Gamma_{ij}|C, S, r_{ij}{=}1) = P(\Gamma_{ij}|\vec{F}_{ij}(\mathbf{x})), \\ \propto \exp\left[-\lambda \sum_{\mathbf{x} \in \Gamma_{ij}} ||\vec{F}_{ij}(\mathbf{x}(t))|| \cdot ||\overrightarrow{\Delta\mathbf{x}}(t)||\right] \end{array} \tag{7}$$

where $\lambda > 0$. Note that the least action, given by (6), will have the highest likelihood in (7). But other hypothetical trajectories in the vector field may also get non-zero likelihoods. When a_i is far away from $\mathbf{s}_j$, the total energy needed to cover that trajectory is bound to be large, and consequently uncertainty about a_i's trajectory is large. Conversely, as a_i gets closer to $\mathbf{s}_j$, uncertainty about the trajectory reduces. Thus, (7) corresponds with our intuition about stochasticity of people's motions. We maintain the probabilities for all possible $r_{ij}, j \in S$.

Video Appearance Features. We are also find useful to model appearance of walkable surfaces as $P(I|C) = \prod_{\mathbf{x} \in \Lambda} P(\phi(\mathbf{x})|c(\mathbf{x}){=}1)$, where $\phi(\mathbf{x})$ is a feature vector consisting of: i) RGB color at $\mathbf{x}$, and ii) Binary indicator if $\mathbf{x}$ belongs to the ground surface of the scene. $P(\phi(\mathbf{x})|c(\mathbf{x}){=}1)$ is specified as a two-component Gaussian mixture model, with parameters ψ. ψ and estimated on our given (single) video with latent $c(\mathbf{x})$, not using training data.

The Probabilistic Model. Given a video, observable random variables include a set of appearance features, I, and a set of partially observed, noisy human trajectories $\Gamma^{(0)}$. Our objective is to infer the latent variables $W = \{C, S, R, \Gamma\}$ by maximizing the posterior distribution of W

$$P(W|\Gamma^{(0)}, I) \propto P(C, S, R) P(\Gamma, \Gamma^{(0)}, I|C, S, R), \tag{8}$$

$$\begin{array}{rcl} P(C, S, R) & = & P(C) P(S|C) P(R|S), \\ P(\Gamma, \Gamma^{(0)}, I|C, S, R) & = & P(\Gamma, \Gamma^{(0)}|C, S, R) P(I|C), \\ P(\Gamma, \Gamma^{(0)}|C, S, R) & = & \prod_{i=1}^{M} \sum_{j=1}^{N} P(\Gamma_{ij}|C, S, r_{ij}{=}1). \end{array} \tag{9}$$

The bottom line of (9) sums all partially observed trajectories $\Gamma_{ij}^{(0)}$, and predicted trajectories Γ_{ij}. We use the same likelihood (7) for $P(\Gamma_{ij}|\cdot)$ and $P(\Gamma_{ij}^{(0)}|\cdot)$.

5. Inference

Given $\{I, \Gamma^{(0)}\}$, we infer $W = \{C, S, R, \Gamma\}$ – namely, we estimate the constraint map, the number and layout of dark matter, hidden human intents, and predict human full trajectories until they reach their goal destinations in the unobserved video parts. To this end, we use the data-driven MCMC process [13, 23], which provides theoretical guarantees of convergence to the optimal solution. In each step, MCMC probabilistically samples C, S, and R. This identifies $\{\vec{F}_{ij}(\mathbf{x})\}$ and the Dijkstra trajectories, which are then used for proposing new C, S, and R. Our MCMC inference is illustrated in Figures 2–3.

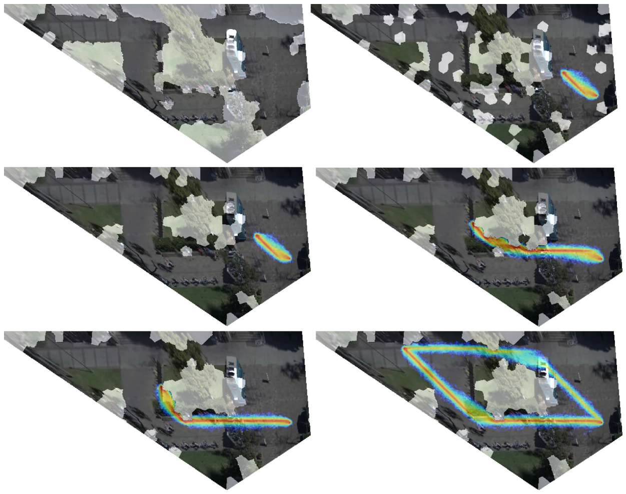

Figure 2. Top view of the scene from Fig. 1 with the overlaid illustration of the MCMC inference. The rows show in raster scan the progression of proposals of the constraint map C (the white regions indicate obstacles), sources S, relationships R, and trajectory estimates (color indicates $P(\Gamma_{ij}|C,S,R)$) of the same person considered in Fig. 1. In the last iteration (bottom right), MCMC estimates that the person's goal is to approach the top-left of the scene, and finds two equally likely trajectories to this goal.

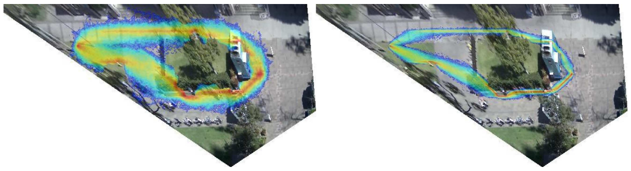

Figure 3. Top view of the scene from Fig. 1 with the overlaid trajectory predictions of a person who starts at the top-left of the scene, and wants to reach the dark matter in the middle-right of the scene (the food truck). A magnitude of difference in parameters $\lambda = 0.2$ (*on the left*) and $\lambda = 1$ (*on the right*) of the likelihood $P(\Gamma_{ij}|C,S,R)$ gives similar trajectory predictions. The predictions are getting more certain as the person comes closer to the goal. Warmer colors represent higher probability.

For stochastic proposals, we use Metropolis-Hastings (MH) reversible jumps. Each jump proposes a new solution Y'=$\{C',S',R'\}$. The decision to discard the current solution, Y=$\{C,S,R\}$, and accept Y' is made based on the acceptance rate, $\alpha = \min\left(1, \frac{Q(Y\to Y')}{Q(Y'\to Y)}\frac{P(Y'|\Gamma^{(0)},I)}{P(Y|\Gamma^{(0)},I)}\right)$ where the proposal distribution is defined as $Q(Y\to Y') = Q(C\to C')Q(S\to S')Q(R\to R')$ and the posterior distribution $P(Y|\Gamma^{(0)},I) \propto P(C,S,R)P(\Gamma,\Gamma^{(0)},I|C,S,R)$ is given by (8) and (9). If α is larger than a number uniformly sampled from $[0,1]$, the jump to Y' is accepted.

The initial C is proposed by setting $c(\mathbf{x}) = 1$ at all locations covered by $\Gamma^{(0)}$, and randomly setting $c(\mathbf{x}) = -1$ or $c(\mathbf{x}) = 1$ for all other locations. The initial number N of sources in S is probabilistically sampled from the Poisson distribution of (3), while their layout is estimated as N most frequent stopping locations in $\Gamma^{(0)}$. Given $\Gamma^{(0)}$ and S, we probabilistically sample the initial R using the multinomial distribution in (4). In the next iteration, the jump step sequentially proposes C', S', and R'.

The Proposal of C' randomly chooses $\mathbf{x} \in \Lambda$, and reverses its polarity, $c'(\mathbf{x}) = -1 \cdot c(\mathbf{x})$. The proposal distribution $Q(C\to C') = Q(c'(\mathbf{x}))$ is data-driven. $Q(c'(\mathbf{x}) = 1)$ is defined as the normalized average speed of people observed at $\mathbf{x}$, and $Q(c'(\mathbf{x}) = -1) = 1 - Q(c'(\mathbf{x}) = 1)$.

The Proposal of S' includes the "death" and "birth" jumps. The birth jump randomly chooses $\mathbf{x} \in \Lambda_C$ and adds a new source $\mathbf{s}_{N+1} = (\boldsymbol{\mu}_{N+1}, \Sigma_{N+1})$ to S, resulting in $S' = S \cup \{\mathbf{s}_{N+1}\}$, where $\boldsymbol{\mu}_{N+1} = \mathbf{x}$ and $\Sigma_{N+1} = \text{diag}(n^2, n^2)$, where n is the scene size (in pixels). The death jump randomly chooses an existing source $\mathbf{s}_j \in S$ and removes it from S, resulting in $S' = S \setminus \{\mathbf{s}_j\}$. The ratio of the proposal distributions is specified as $\frac{Q(S\to S')}{Q(S'\to S)} = 1$, indicating no preference to either 'death" or "birth" jumps. That is, the proposal of S' is exclusively governed by the Poisson prior of (3), and trajectory likelihoods $P(\Gamma_{ij}|C',S',R)$, given by (7), when computing the acceptance rate α.

The Proposal of R' randomly chooses one person $a_i \in A$ with goal $\mathbf{s}_j$, and randomly changes a_i's goal to $\mathbf{s}_k \in S$. This changes the corresponding relationships $r_{ij}, r_{ik} \in R$, resulting in R'. The ratio of the proposal distributions is $\frac{Q(R\to R')}{Q(R'\to R)} = 1$. This means that the proposal of R' is exclusively governed by the multinomial prior $P(R'|S')$, given by (4), and trajectory likelihoods $P(\Gamma_{ij}|C',S',R')$, given by (7), when computing the acceptance rate α.

From the accepted jumps C', S' and R', we can readily update the force fields $\{\vec{F}'_{ij}\}$, given by (5), and then compute the Dijkstra paths of every person $\{\Gamma'_{ij}\}$ as in (6).

6. Experiments

Our method is evaluated on toy examples and 4 real outdoor scenes. We present three types of results: (a) localization of "dark matter" S, (b) estimation of human intents R, and (c) trajectory prediction Γ. Annotating ground truth of constraint map C in a scene is difficult, since human annotators provide inconsistent subjective estimates. Therefore, we do not estimate our inference of C. Our evaluation advances that of related work [12], which focuses only on detecting " exits" and " vehicles" in the scene, and predicting human trajectories. Note that a comparison with existing approaches to object detection would be unfair, since we only have the video as our input and do not have access to annotated examples of the objects, as most appearance-based methods for object recognition.

Metrics. Negative Log-Likelihood (**NLL**) and Modified Hausdorff Distance (**MHD**) are measured to evaluate trajectory prediction. $P(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)})$ is given by (7), NLL of a true trajectory $\mathbf{X} = \{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(T)}\}$ is defined as

$$\text{NLL}_P(\mathbf{X}) = -\frac{1}{T-1}\sum_{t=1}^{T-1}\log(P(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)})) \qquad (10)$$

MHD between true trajectory $\mathbf{X}$ and our sampled trajectory $\mathbf{Y} = \{\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(T)}\}$ is defined as

$$\begin{aligned} \text{MHD}(\mathbf{X}, \mathbf{Y}) &= \max(d(\mathbf{X}, \mathbf{Y}), d(\mathbf{Y}, \mathbf{X})) \\ d(\mathbf{X}, \mathbf{Y}) &= \frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}\in\mathbf{X}}\min_{\mathbf{y}\in\mathbf{Y}}||\mathbf{x}-\mathbf{y}|| \end{aligned} \qquad (11)$$

We present the average MHD between the true trajectory and our 5000 trajectory prediction samples. For evaluating detection of S, we use the standard overlap criterion of our detection and ground-truth bounding box around the functional object of interest. When the ratio of intersection over union of our detection and ground-truth bounding box is larger than 0.5, we deem the detection true positive. For evaluation of predicting human intents R, we allow our inference access to an initial part of the video footage, in which R is not observable, and then compare our results with ground-truth outcomes of R in the remaining (unobserved) video parts.

Baselines. Our baseline for estimating S is an initial guess of "dark matter" based on partial observations $\{\Gamma^{(0)}, I\}$, before our DDMCMC inference. This baseline declares every location in the scene as "dark matter" at which the observed people trajectories in $\Gamma^{(0)}$ ended, and people stayed still at that location longer than 5sec before changing their trajectory. The baseline of estimating R is a greedy move (GM) algorithm $P(r_{ij}|\{\Gamma_i^{(0,\cdots,t)}\}) \propto \exp\{\tau(||\mathbf{x}_j - \Gamma_i^{(t)}|| - ||\mathbf{x}_j - \Gamma_i^{(0)}||)\}$. We also use the following three naive methods as baselines. (1) Shortest path (SP) estimates the trajectory as a straight line, disregarding obstacles in the scene. (2) Random Walk (RW). (3) Lagrangian Physical Move (PM) under the sum of all forces from multiple fields, $\vec{F}_{\text{classic}}(\mathbf{x})$, as defined in Sec. 4, as by the classical LM.

Comparison with Related Approaches. We are not aware of prior work on estimating S and R in the scene without access to training labels of objects. So we compare only with the state-of-the-art method for trajectory prediction [12].

Parameters. In our setting, the first 50% of a video is observed, and human trajectories in the entire video is to be predicted. We use the following model parameters: $\beta = .05$, $\lambda = .5$, $\rho = .95$. From our experiments, varying these parameters in intervals $\beta \in [.01, .1]$, $\lambda \in [.1, 1]$, and $\rho \in [.85, .98]$ does not change our results, suggesting that we are relatively insensitive to the specific choices of β, λ, ρ over certain intervals. η is known. θ and ψ are fitted from observed data.

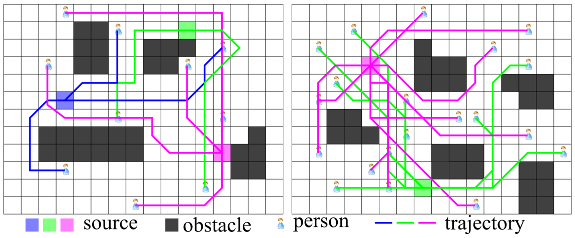

Figure 4. Two samples of toy examples.

$\|S\|$	S, R				NLL			
	10	20	50	100	10	20	50	100
2	0.95	0.97	0.96	0.96	1.35	1.28	1.17	1.18
3	0.87	0.90	0.94	0.94	1.51	1.47	1.35	1.29
5	0.63	0.78	0.89	0.86	1.74	1.59	1.36	1.37
8	0.43	0.55	0.73	0.76	1.97	1.92	1.67	1.54

Table 1. Results of toy example. Left is accuracy of $S\&R$, it's counted correct only if both S and R are correct. Right is NLL. Second row is number of agents $|A|$, first column is number of sources $|S|$.

Dataset	$\|S\|$	Source Name
① Courtyard	19	bench/chair,food truck, bldg, vending machine, trash can, exit
② SQ1	15	bench/chair, trash can, bldg, exit
③ SQ2	22	bench/chair, trash can, bldg, exit
④ VIRAT	17	vehicle, exit

Table 2. Summary for datasets

6.1. Toy example

The toy example allows us to methodologically test our approach with respect to each dimension of the scene complexity, while fixing the other dimensions. The scene complexity is defined in terms of the number of agents in the scene and the number of sources. These parameters are varied to synthesize the toy artificial scenes. The toy example is in a rectangle random layout, the ratio of obstacle pixels over all pixels is about 15%, the ratio of observed part of trajectories is about 50%. We vary $|S|$ and $|A|$, and we have 3 repetitions for each setting. Tab. 1 shows that our approach can handle large variations in each dimension of the scene complexity.

6.2. Real scenes

Datasets. We use 4 different real scenes for evaluation: ① Courtyard dataset [3]; and our new video sequences of two squares ② SQ1 and ③ SQ2 annotated by VATIC [25]; ④ VIRAT ground dataset [16]. SQ1 is 20min, 800 × 450, 15 fps. SQ2 is 20min, 2016 × 1532, 12 fps. We use the same scene A of VIRAT as in [12]. We allow initial (partial) observation of 50% of the video footage, which for example gives about 300 trajectories in ①.

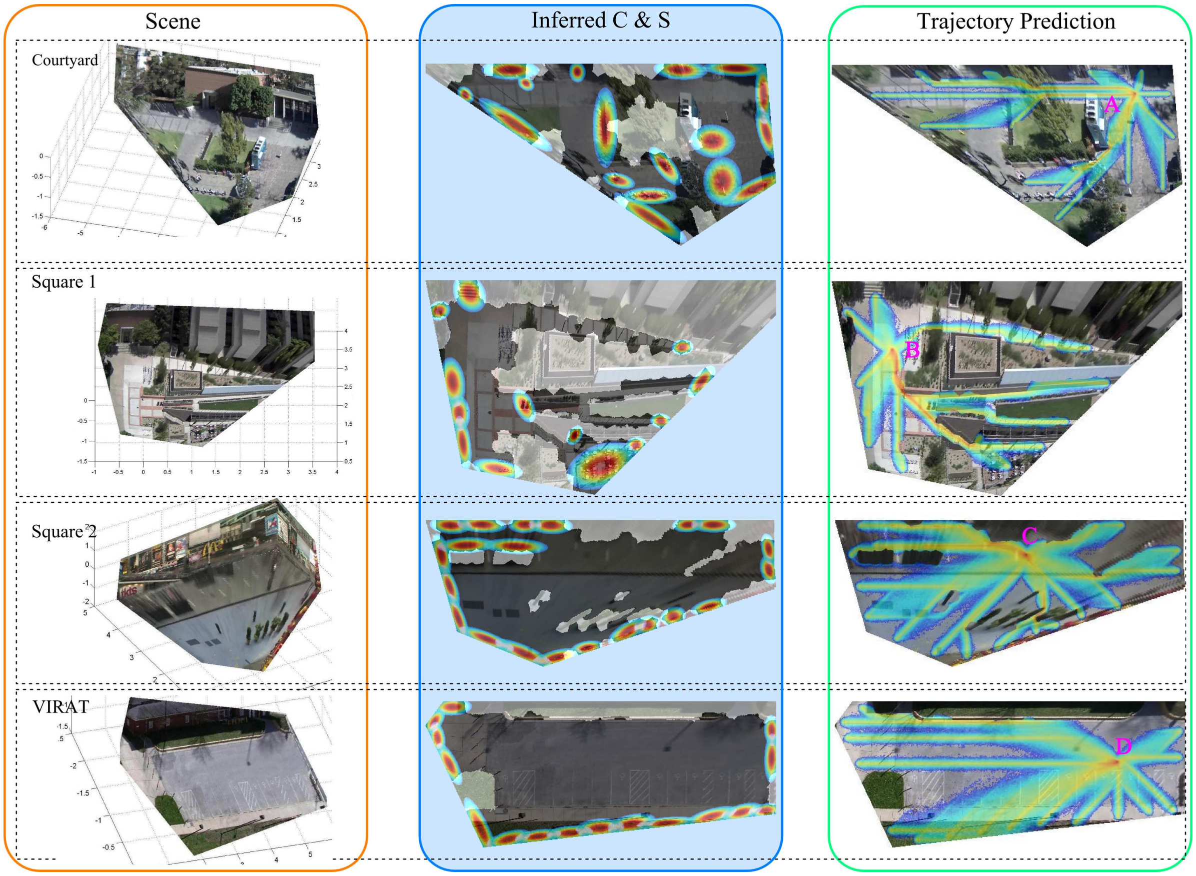

Figure 5. Qualitative experiment results for 4 scenes. Each row is one scene. The 1st column is the reconstructed 3D surfaces of each scene. The 2nd column is the estimated layout of obstacles (the white masks) and dark matter (the Gaussians). The 3rd column is an example of trajectory prediction by sampling, we predict the future trajectory for a particular agent at some position (A, B, C, D) in the scene toward each potential source in S, the warm and cold color represent high and low probability of visiting that position respectively.

Dataset	S		R		NLL			MHD			
	Our	Initial	Our	GM	Our	[12]	RW	Our	RW	SP	PM
①	**0.89**	0.23	**0.52**	0.31	**1.635**	-	2.197	**17.4**	243.1	43.2	207.5
②	**0.87**	0.37	**0.65**	0.53	**1.459**	-	2.197	**11.6**	262.1	39.4	237.9
③	**0.93**	0.26	**0.49**	0.42	**1.621**	-	2.197	**21.5**	193.8	27.9	154.2
④	**0.95**	0.25	**0.57**	0.46	**1.476**	1.594	2.197	**16.7**	165.4	21.6	122.3

①Courtyard	45%	40%
S	0.85	0.79
R	0.47	0.41
NLL	1.682	1.753
MHD	21.7	28.1

Table 3. Left: Results of 4 real scenes. The results show that our approach outperform the baselines. The accuracy of S verifies that these dark matter can be recognized through human activities. Intent prediction R by our method is better than GM, and the accuracy is higher when S is smaller. The trajectory prediction (NLL and MHD) is more accurate is constrained scene (① ②) than free scenes (③④). Right: Results of scene ①Courtyard with different observed ratio. The performance downgrades gracefully with smaller observed ratio.

Results. The qualitative results for real scenes are shown in Fig. 5 and the quantitative evaluation is presented in Tab. 3. As can be seen: (1) We are relatively insensitive to the specific choice of model parameters. (2) We handle challenging scenes with arbitrary layouts of dark matter, both in the middle of the scene and at its boundaries. From Tab. 3, the comparison with the baselines demonstrates that the initial guess of sources based on partial observations gives very noisy results. These noisy results are significantly improved in our DD-MCMC inference. Also, our method is a slightly better than the baseline GM if there are a few obstacles in the middle of the scene. But we get a huge performance improvement over GM if there are complicated obstacles in the scene. This shows that our global plan based relation prediction is better than GM. We are also superior to the random walk. The baselines RW and PM produce bad trajectory prediction. While SP yields good results for scenes with a few obstacles, it is brittle for more complex scenes which we successfully handle. When the size of S is large (*e.g.*, many exists from the scene), our estimation of human goals may not be exactly correct. However, in all these error cases, the goal that we estimate is not spatially far away from the true goal. Also, in these cases, the predicted trajectories are also not far away from the true trajectories measured by MHD and NLL. Our performance downgrades gracefully with the reduced observation time.

We outperform the state of the art [12]. Note that the MHD absolute values produced by our approach and [12] are not comparable, because this metric is pixel based and depends on the resolution of reconstructed 3D surface.

Our results show that our method successfully addresses surveillance scenes of various complexities.

7. Conclusion

We have addressed a new problem, that of localizing functional objects in surveillance videos without using training examples of objects. Instead of appearance features, human behavior is analyzed for identifying the functional map of the scene. We have extended the classical Lagrangian mechanics to model the scene as a physical system wherein: i) functional objects exert attraction forces on people's motions, and ii) people are not inanimate particles but agents who can have intents to approach particular functional objects. Given a small excerpt from the video, our approach estimates the constraint map of non-walkable locations in the scene, the number and layout of functional objects, and human intents, as well as predicts human trajectories in the unobserved parts of the video footage. For evaluation we have used the benchmark VIRAT and UCLA Courtyard datasets, as well as our two 20min videos of public squares.

Acknowledgements

This research has been sponsored in part by grants DARPA MSEE FA 8650-11-1-7149, ONR MURI N00014-10-1-0933, NSF IIS 1018751, and NSF IIS 1018490.

References

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

[2] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *EECV*, 2008. 3

[3] 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. 2, 6

[4] C. L. Baker, R. Saxe, and J. B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 2009. 3

[5] J. Barraquand, B. Langlois, and J.-C. Latombe. Numerical potential field techniques for robot path planning. *TSMC*, 1992. 3

[6] J. Gall, A. Fossati, and L. V. Gool. Functional categorization of objects using real-time markerless motion capture. In *CVPR*, 2011. 2

[7] H. Gong, J. Sim, M. Likhachev, and J. Shi. Multi-hypothesis motion planning for visual object tracking. In *ICCV*, 2011. 3

[8] H. Grabner, J. Gall, and L. V. Gool. What makes a chair a chair ? In *CVPR*, 2011. 2

[9] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: using spatial and functional compatibility for recognition. *TPAMI*, 2009. 2

[10] M. Hoai and F. De la Torre. Max-margin early event detectors. In *CVPR*, 2012. 3

[11] K. Kim, M. Grundmann, A. Shamir, I. Matthews, J. Hodgins, and I. Essa. Motion fields to predict play evolution in dynamic sport scenes. In *CVPR*, 2010. 3

[12] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *ECCV*, 2012. 3, 5, 6, 7, 8

[13] J. Kwon and K. M. Lee. Wang-Landau monte carlo-based tracking methods for abrupt motions. *TPAMI*, 2013. 2, 4

[14] K. H. Lee, M. G. Choi, Q. Hong, and J. Lee. Group behavior from video : A data-driven approach to crowd simulation. In *SCA*, 2007. 3

[15] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In *Eurographics*, 2007. 3

[16] S. Oh et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, 2011. 2, 6

[17] M. Pei, Y. Jia, and S.-C. Zhu. Parsing video events with goal inference and intent prediction. In *ICCV*, 2011. 3

[18] S. Pellegrini, J. Gall, L. Sigal, and L. V. Gool. Destination flow for crowd simulation. In *ECCV*, 2012. 3

[19] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 2011. 2

[20] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *ICCV*, 2011. 3

[21] W. Shao and D. Terzopoulos. Autonomous pedestrians. In *SCA*, 2005. 3

[22] B. Solmaz, B. E. Moore, and M. Shah. Identifying behaviors in crowd scenes using stability analysis for Dynamical Systems. *TPAMI*, 2012. 3

[23] Z. Tu and S.-C. Zhu. Image segmentation by data-driven markov chain monte carlo. *TPAMI*, 2002. 2, 4

[24] M. W. Turek, A. Hoogs, and R. Collins. Unsupervised learning of functional categories in video scenes. In *ECCV*, 2010. 2

[25] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *IJCV*, 2013. 6

[26] Y. Zhao and S.-C. Zhu. Image parsing via stochastic scene grammar. In *NIPS*, 2011. 2

[27] B. Zhou, X. Wang, and X. Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *CVPR*, 2011. 3

[28] B. Zhou, X. Wang, and X. Tang. Understanding collective crowd behaviors: Learning a Mixture model of Dynamic pedestrian-Agents. In *CVPR*, 2012. 3