2020 American Control Conference
Denver, CO, USA, July 1-3, 2020

Oriented Pedestrian Social Interaction Modeling and Inference

Junyi Dong, Pingping Zhu, and Silvia Ferrari

Abstract— **In order to drive and operate safely around humans, future autonomous vehicles will be expected to perceive visual scenes and predict human behaviors beyond explicit visual features. Inferring human interactions, for example, plays an indispensable role in predicting pedestrian trajectories, because social actions such as walking together, gathering, holding hands, and talking, influence where and how people move relative to each other and their environment. Existing methods for semantic action recognition and labeling provide inputs that, while useful to human operators, cannot be used to improve predictions made by autonomous vehicles. This paper presents a graphical model approach for jointly inferring pedestrian interactions from short video clips over time. New Markov random field algorithms are presented for modeling social interactions probabilistically using spatial and temporal observations obtained over short video clips, at a time scale useful for making real-time decisions such as collision avoidance. Experiments conducted using real-world pedestrian streaming videos show that the average interaction-inference accuracy of the proposed approach is approximately 94.6%.**

I. Introduction

Autonomous vehicles and smart environments will soon require the ability to predict human behaviors and trajectories with high accuracy and well into the future [1], [2], [3]. Existing tracking algorithms can fall short of predicting human actions, often deemed and modeled as random, because human behaviors cannot be captured by kinodynamic differential equation models applicable to robots and vehicles. Human decisions and behaviors, such as jaywalking versus waiting for the light to turn green or looking for a crossroad, are largely driven by factors, such as emotions, social interactions, and internal thoughts, that are not readily extracted from video or sensor data. This paper seeks to obtain a mathematical model of social interactions and familiarity from streaming video of pedestrians, such that significant interrelationships can be inferred from a scene, similarly to cognitive processes of inference that allow a human driver to predict people trajectories based on visual features alone.

Most of the existing human tracking and prediction literature treats pedestrians as a group of mutually independent individuals, without taking into account social interactions. Yet, statistical video analysis reveals that the trajectories of individuals who are interacting socially while walking are highly correlated (Section V-B). Inferring human interactions from video is a challenging problem. To date, methods that account for human interactions in order to improve trajectory prediction and robot navigation can be distinguished into empirical, clustering, and inference based [4], [5], [6]. Empirical methods rely on a Euclidean-distance threshold obtained from social studies and experiments in order to determine whether two or more people are interacting. The empirical approach performs poorly in complex, pedestrian-rich environments because, as shown in Section V-B, proximity measures vary greatly and dynamically as people move about. Clustering methods, such as the dominant set (DS) algorithm, develop graph representations of human interactions by representing each person as a node and pairwise interactions as arcs [5]. Interacting people are clustered into maximum cliques, typically resulting in a dense interaction structure. The inference method presented in [6] represents interactions by binary random variables in an agent-based model derived from first principles and observations of the pedestrians' position, speed, preferred speed, and chosen destination.

This paper seeks to combine the advantages of both clustering and inference methods by modeling and inferring pedestrian interactions using the probabilistic graphical models known as Markov random fields (MRFs). A new MRF model is presented such that nodes can be used to represent pedestrian bounding boxes extracted from video, and probabilistic arcs can be used to represent pedestrian interactions. In this case, the graph structure is inferred from data, by determining the maximum *a-posteriori* (MAP) estimate of the graph arc set. The exponential complexity of the MAP inference problem is addressed by designing an energy function that allows to convert the inference problem into an integer linear program (ILP). Furthermore, the energy function is constructed to encode discriminant interaction features, such as position, speed, and orientation, and such that its parameters can be learned from a small labeled training set, using a structural support vector machine (SSVM) algorithm [7].

When compared to existing methods, the advantages of the MRF approach presented in this paper are threefold. Firstly, the problem of inferring human interactions is formulated from the perspective of jointly inferring the arcs of an undirected graphical model, whereby the symmetric property of interaction is automatically guaranteed. Secondly, a flexible MRF inference algorithm is developed by sharing parameters among all arcs, such that the model can be applied to scenes with an arbitrary number of pedestrians. Thirdly, the MRF approach can be generalized to other scene interpretation problems and, potentially, used to infer other hidden variables associated with human labels and interactions. The proposed MRF approach is shown to achieve an average interaction inference accuracy of approximately 95%, when tested across very different videos and outdoor scenarios.

Junyi Dong (jd979@cornell.edu), Pingping Zhu (pz224@cornell.edu), and Silvia Ferrari (sf375@cornell.edu) are with the Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY, US.

978-1-5386-8266-1/$31.00 ©2020 AACC

II. Problem formulation

The prediction of people trajectories and behaviors is relevant to a broad range of applications including but not limited to autonomous driving, autonomous robots, and human-machine interactions. Modern computer vision algorithms are able to extract and track people bounding boxes [8], and, using past measurements, to estimate future trajectories by methods such as Kalman filters, particle filters, or Bayesian nonparametric models [9]. However, people trajectories and behaviors are also highly influenced by mutual interactions and relationships that cannot be readily extracted from video or camera frames. Therefore, this paper seeks to develop a model of human interactions based on video recordings obtained over a finite time window, without prior knowledge.

Consider a video of pedestrians, $\mathcal{V}$, obtained over a finite and fixed time window, $[t_0,\ t_f]$, using a fixed camera in an outdoor environment, with no access to verbal communications. The video is comprised of many camera frames obtained at discrete moments in time in $[t_0,\ t_f]$. Therefore, each video frame is represented by an image matrix of $(m \times n)$ pixel intensities, denoted by $I_l \in \mathbb{R}^{m\times n}$, where m and n are known camera parameters, and l indicates the frame index. A labeled bounding box can be obtained for each pedestrian in a frame I_l using a convolutional neural network (CNN), as shown in [8], [10].

The number of pedestrians in the camera field-of-view (FOV) changes over time and can be obtained from the number of bounding boxes in each frame. Therefore, after a video with M_f total frames is obtained, i.e.,

$$\mathcal{V} = \{I_l \mid I_l \in \mathbb{R}^{m\times n},\ l = 1,\dots,M_f\} \tag{1}$$

it can be partitioned into short, consecutive, non-overlapping video clips that each contain a fixed number of pedestrians, as follows. Let V_k denote the k^{th} video clip in $\mathcal{V}$ containing N_k pedestrians detected from M_k consecutive frames with frame-index set T_k, or

$$V_k = \{I_l \mid I_l \in \mathbb{R}^{m\times n},\ l \in T_k\}, \quad k = 1,\dots,f \tag{2}$$

where f indicates the total number of video clips. Then, $\mathcal{V}$ can be partitioned into a set of non-overlapping short video clips, such that,

$$\mathcal{V} = \cup_{k=1}^{f} V_k, \text{ and } V_k \cap V_{k'} = \emptyset, \text{ if } k \neq k' \tag{3}$$

where $I_1 \in V_1$ is the first frame obtained at time t_0, and $I_{M_f} \in V_f$ is the last frame obtained at time t_f.

The goal of this paper is to model and infer pedestrian interactions dynamically, during each video clip, in order to aid in the prediction of their future trajectories and behaviors. For simplicity, each pedestrian is assumed to interact with at most one other pedestrian. A lonesome pedestrian is referred to as singleton hereon in the paper. Although these assumptions are met in most scenarios [11], future work will consider larger group interactions and more complex behaviors. The goal of the model is to infer paired pedestrian interactions whereby a symmetric relationship is induced by social acquaintance or familiarity and is recorded by a fixed stationary camera in an outdoor environment.

III. MRF Model of Pedestrian Interactions

Markov random fields or MRFs are undirected probabilistic graphical models defined over a set of discrete or continuous random variables that may be hidden or observable [12], [13]. In traditional MRFs, each node represents a random variable and the arc set, or graph structure, represents a factorization of the joint MRF probability that is learned from data [14]. Typically, the arc set is pre-defined to represent a regular structure such as a uniform grid [15] or a fully connected graph [16]. In contrast, this paper presents a new MRF model that can be used to model and infer pedestrian relationships from streaming video data by representing pedestrian labels as nodes and their interrelationships as arcs. Then, the MRF structure is inferred probabilistically in order to determine the arc configuration with the maximum posterior probability (MAP) by minimizing an energy function learned from data.

A. MRF Structure

Unlike previous MRF methods for computer vision in which the graph structure represented the most probable image segmentation [13], [15], this paper develops an MRF approach for modeling and inferring hidden pedestrian relationships from multiple, consecutive video frames. The MRF pedestrian model is dynamically constructed with every video clip V_k, $k = 1,\dots,f$, with few offline training data and based on video frames acquired from the fixed camera FOV.

Let $\mathcal{N}_k = \{1,\dots,N_k\}$, $N_k \in \mathbb{N}^+$, denote the index set of pedestrian extracted from V_k via CNN, where N_k is obtained by counting the bounding boxes' labels. Every pedestrian bounding box extracted from V_k is represented by an MRF node labeled by the bounding box index $i \in \mathcal{N}_k$. As a result, $\mathcal{N}_k$ defines the full set of MRF nodes, which can be assumed observable and known from the video clip V_k. The set of undirected arcs $\mathcal{E}_k = \{(i,j) \mid i,j \in \mathcal{N}_k\}$ represents the pedestrian interactions in the scene, such that an arc (i,j) is placed between the node representations of bounding boxes i and j if the corresponding pedestrians are believed to interact significantly with one another in V_k. The singleton case of a lonesome pedestrian with bounding box label i is represented by an arc (i,i) connecting node i to itself. Both arc representations are illustrated by the orange line in Fig. 1. Then, the MRF pedestrian model structure to be learned from a video clip V_k is given by the pair $(\mathcal{N}_k, \mathcal{E}_k)$.

Unlike pedestrian bounding boxes, which can be extracted from V_k using CNNs or other computer vision algorithms, pedestrian interactions due to social acquaintance or familiarity come in many different forms and, typically, cannot be readily extracted from video frames. This is because similar instantaneous pedestrian positions, poses, and behaviors may be induced by social interactions or by chance. While training a CNN to recognize these implicit interactions from streaming video is a possible solution [17], [18], it requires training on very large data sets and may lead to predictions that are not necessarily robust to the broad range of social

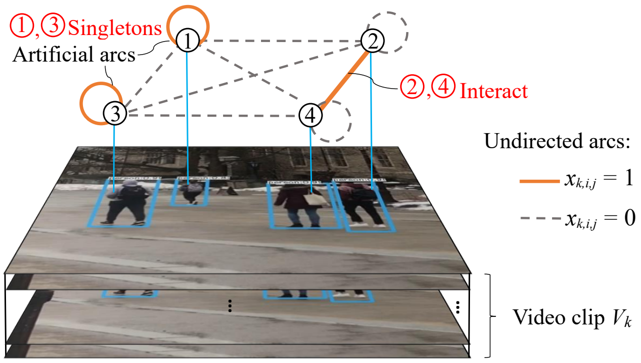


Fig. 1. Interaction model of four pedestrians based on corresponding (unlabeled) bounding boxes shown in blue.

situations and behaviors exhibited by pedestrians in different contexts and settings.

Because the pedestrian interactions are to be inferred from short video clips, in this paper, the MRF structure is inferred from data, by assigning a random binary variable $X_{k,i,j}$ to each MRF arc $(i,j) \in \mathcal{E}_k$, such that its value $x_{k,i,j}$ equals one when a symmetric interaction exists between pedestrians labeled by i and j, and equals zero when an interaction does not exist, i.e., $x_{k,i,j} \in \mathcal{L}$, where $\mathcal{L} = \{0,1\}$. Then, all random variables can be organized into a matrix,

$$\mathbf{X}_k \triangleq (X_{k,i,j})_{N_k \times N_k} \qquad k = 1, \ldots, f \tag{4}$$

that is to be inferred from video data, and any arc set $\mathcal{E}_k$ is a possible realization (or instantiation) of $\mathbf{X}_k$ that has range $\mathcal{L}^{N_k^2}$.

For each frame I_l $(l \in T_k)$ in video clip V_k, the i^{th} pedestrian's position, $\mathbf{p}_{k,i,l} \in \mathbb{R}^{2\times 1}$, speed $v_{k,i,l} \in \mathbb{R}^+$, and heading $\theta_{k,i,l} \in [0, \ 2\pi)$ can be extracted. The 2D position and velocity vectors of each pedestrian are measured relative to the FOV, so as not to require knowledge of the camera position and orientation in inertial frame. Then, the pedestrian speed and heading are obtained from the norm and orientation of the velocity vector with respect to the horizontal FOV direction, respectively. Next, organize all measurements into a frame observation vector,

$$\mathbf{z}_{k,i,l} \triangleq [\mathbf{p}_{k,i,l}^T \quad v_{k,i,l} \quad \theta_{k,i,l}]^T \tag{5}$$

Then, the sequence of consecutive observations extracted over the entire video clip V_k can be organized into a video observation vector,

$$\mathbf{z}_{k,i} = [\mathbf{z}_{k,i,1}^T \ \ldots \ \mathbf{z}_{k,i,|T_k|}^T]^T \tag{6}$$

obtained by stacking all (column) observation vectors for pedestrian i, obtained from the video clip V_k, where $|T_k|$ denotes the cardinality of index set T_k. Finally, the observation vectors obtained from all pedestrians in video clip V_k are organized into an $(4|T_k| \times N_k)$ video observation matrix,

$$\mathbf{Z}_k = [\mathbf{z}_{k,1} \ \ldots \ \mathbf{z}_{k,N_k}] \tag{7}$$

Then, the goal of the MRF structural learning algorithm presented in Section IV is to infer the optimal set of arcs, $\mathcal{E}_k^*$, from the video observation matrix, $\mathbf{Z}_k$.

B. MRF Energy Function

The structural inference problem can be transcribed into tractable optimization problem by designing an energy function that can be learned from a small set of discriminative features governing pedestrian interactions. Although spatial proximity has been used as an interaction feature in several published works [4], [5], [6], it alone is not discriminative enough in dynamic scenes populated with pedestrians. For example, spatial proximity would not discriminate among a scenario with two interacting pedestrians walking side by side (Fig. 2.a) and scenarios with two non-interacting pedestrians passing by (Fig. 2.b) or crossing paths (Fig. 2.c).

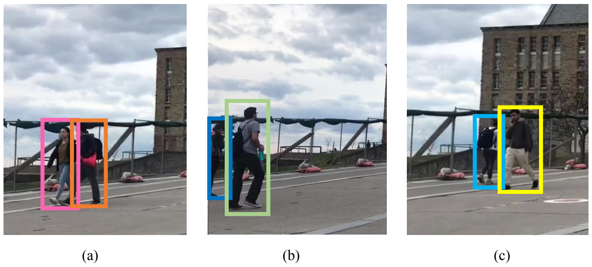

(a)
(b)
(c)

Fig. 2. Examples of pedestrians walking and interacting (a), passing by and non-interacting (b), crossing paths and non-interacting (c), and corresponding bounding boxes (labeled by color).

Therefore, this paper presents a new discriminative feature representation for any node pair (i,j) that is based on proximity (relative position), as well as relative velocity and relative heading, i.e.,

$$\phi(\mathbf{z}_{k,i}, \mathbf{z}_{k,j}) \triangleq \frac{1}{|T_k|}\begin{bmatrix} \sum_{l \in T_k} \|\mathbf{p}_{k,i,l} - \mathbf{p}_{k,j,l}\| \\ \sum_{l \in T_k} \|v_{k,i,l} - v_{k,j,l}\| \\ \sum_{l \in T_k} \|\theta_{k,i,l} - \theta_{k,j,l}\| \end{bmatrix}, \quad \forall i,j \in \mathcal{N}_k, i \neq j \tag{8}$$

where $\|\cdot\|$ is the Euclidean norm. The observations are averaged over the video clip length in order to obtain a robust and representative estimate for the video clip V_k. To subsume the singleton case into a unified framework, we let $\phi(\mathbf{z}_{k,i}, \mathbf{z}_{k,i}) \triangleq [d_p \quad d_v \quad d_\theta]^T$, where $d_p, d_v, d_\theta \in \mathbb{R}^+$ are user-defined hyper-parameters that can be interpreted as the counterparts of proximity, relative velocity, and relative heading in (8), respectively.

After the hyper-parameters are tuned based on the video data (Section IV-A), the generalized feature representation can be written in a compact form,

$$\boldsymbol{\phi}(\mathbf{z}_{k,i}, \mathbf{z}_{k,j}) \triangleq \frac{1}{|T_k|} \begin{bmatrix} \sum_{l \in T_k} \|\mathbf{p}_{k,i,l} - \mathbf{p}_{k,j,l}\| \, \delta(i \neq j) + d_p \ \delta(i = j) \\ \sum_{l \in T_k} \|v_{k,i,l} - v_{k,j,l}\| \, \delta(i \neq j) + d_v \ \delta(i = j) \\ \sum_{l \in T_k} \|\theta_{k,i,l} - \theta_{k,j,l}\| \, \delta(i \neq j) + d_\theta \ \delta(i = j) \end{bmatrix}, \quad \forall i, j \in \mathcal{N}_k \tag{9}$$

where $\delta(\cdot)$ is an indicator function that equals one when the enclosed statement holds true and is zero otherwise.

From the above feature representation, an arc potential function is designed to relate the discriminative features to the random variable $X_{k,i,j}$,

$$\Phi(\mathbf{z}_{k,i}, \mathbf{z}_{k,j}, x_{k,i,j}, \mathbf{w}) = [\mathbf{w}^T \boldsymbol{\phi}(\mathbf{z}_{k,i}, \mathbf{z}_{k,j})] \ x_{k,i,j} \tag{10}$$

where $x_{k,i,j}$ is the realization of $X_{k,i,j}$, and $\mathbf{w}$ is a vector of parameters to be learned from data. Then, using the approach in [12], the MRF energy function can be defined as the sum of arc potentials,

$$E(\mathbf{Z}_k, \mathcal{E}_k) \triangleq \sum_{(i,j) \in \mathcal{E}_k} [\mathbf{w}^T \boldsymbol{\phi}(\mathbf{z}_{k,i}, \mathbf{z}_{k,j})] \ x_{k,i,j} \tag{11}$$

It can be seen that the energy function is constructed as a linear function of the realizations of the random variables $X_{k,i,j}$, when $\mathbf{w}$ is given. Therefore, parameter learning is performed as a first step using a training data set in which the interaction ground truth is available. Subsequently, structural inference is approached as an energy minimization problem that can be solved via Integer Linear Programs (ILPs).

IV. MRF Learning and Inference

By conditioning structural inference on the video observation matrix $\mathbf{Z}_k$, the MRF joint distribution can be written as the normalized negative exponential of an energy function,

$$P(\mathcal{E}_k \mid \mathbf{Z}_k) = \frac{1}{C} \exp\{-E(\mathbf{Z}_k, \mathcal{E}_k)\} \tag{12}$$

where C is a normalization constant [14], and $E(\mathbf{Z}_k, \mathcal{E}_k)$ is defined in (11). Then, inferred arcs of high probability correspond to minima of the energy function. In particular, the optimal labeling $\mathcal{E}_k^*$ that best describes the pedestrian interaction structure is obtained from the MAP estimate of the random variables $\mathbf{X}_k$, or equivalently, by minimizing the energy function,

$$\mathcal{E}_k^* = \arg\max_{\mathcal{E}_k} \{P(\mathcal{E}_k \mid \mathbf{Z}_k)\} = \arg\min_{\mathcal{E}_k} \{E(\mathbf{Z}_k, \mathcal{E}_k)\} \tag{13}$$

subject to the constraints,

$$\sum_{i \in \mathcal{N}_k} x_{k,i,j} = 1, \quad \forall j \in \mathcal{N}_k \tag{14}$$

$$x_{k,i,j} = x_{k,j,i}, \quad \forall (i,j) \in \mathcal{E}_k \tag{15}$$

$$\mathcal{E}_k \in \mathcal{L}^{N_k^2} \tag{16}$$

which guarantee that a pedestrian is either a singleton or interacts with, at most, one other pedestrian, and all pairwise interactions are symmetric.

The next subsections describe the algorithms for learning first the MRF parameters and, then, for inferring the optimal graph structure by minimizing the parameterized function.

A. Parameter Learning via SVM

This subsection presents an approach for learning the energy function parameters from a database of manually annotated ground-truth arcs, denoted by $\breve{\mathcal{E}}_k$, and corresponding video observation matrices, i.e.,

$$\mathcal{D} = \{(\breve{\mathcal{E}}_1, \mathbf{Z}_1), \ldots, (\breve{\mathcal{E}}_f, \mathbf{Z}_f)\} \tag{17}$$

where the total number of video clips, f, may or may not be the same as in the validation database. The learning objective is to find the optimal parameters $\mathbf{w}^*$ for which the true arc set $\breve{\mathcal{E}}_k$ has the lowest energy value than any other set $\mathcal{E}_k \in \mathcal{L}^{N_k^2}$, according to the energy function in (11).

Let ξ_ℓ ($\ell = 1, \ldots, f$) denote f slack variables, and let c denote a constant weight that balances the minimization of the parameter magnitude and the minimization of the energy function [7]. Then, the optimal MRF parameters $\mathbf{w}^*$ can be obtained by solving the constrained optimization problem,

$$\min_{\mathbf{w}, \xi_\ell} \quad \frac{1}{2} \|\mathbf{w}\|^2 + \frac{c}{f} \sum_{\ell=1}^{f} \xi_\ell \tag{18}$$

$$\text{sbj to} \quad E(\mathbf{Z}_k, \mathcal{E}_k) - E(\mathbf{Z}_k, \breve{\mathcal{E}}_k) \geq 1 - \xi_\ell, \quad \forall k, \ell \tag{19}$$

$$\mathbf{w} \geq \mathbf{0}, \ \xi_\ell \geq 0, \quad \forall \ell, \quad \forall \mathcal{E}_k \neq \breve{\mathcal{E}}_k, \quad k = 1, \ldots, f$$

using a structural SVM method [7]. The constraints in (19) ensure that the energy of the true graph structure ($\breve{\mathcal{E}}_k$) is less than that of any other structure by a margin controlled by the slack variables (ξ_ℓ).

B. MRF Structural Inference

Once the optimal energy function parameters are learned from the training database $\mathcal{D}$ (Section IV-A), the energy function $E(\mathbf{Z}_k, \mathcal{E}_k, \mathbf{w}^*)$ can be minimized to infer the optimal pedestrian interactions in a new validation database comprised of a streaming video, $\mathcal{V}$. By combining the energy function in (11) with the structural inference constraints in (14)-(15), the inference problem in (13) is re-formulated as an ILP in $\mathcal{E}_k$ with linear objective and constraints:

$$\min_{\mathcal{E}_k} \quad \sum_{(i,j) \in \mathcal{E}_k} [(\mathbf{w}^*)^T \boldsymbol{\phi}(\mathbf{z}_{k,i}, \mathbf{z}_{k,j})] \ x_{k,i,j} \tag{20}$$

$$\text{sbj to} \quad \sum_{i \in \mathcal{N}_k} x_{k,i,j} = 1, \quad \forall j \in \mathcal{N}_k \tag{21}$$

$$x_{k,i,j} = x_{k,j,i}, \quad \forall (i,j) \in \mathcal{E}_k \tag{22}$$

$$\mathcal{E}_k \in \mathcal{L}^{N_k^2} \tag{23}$$

The ILP structure constraints are crucial to enabling the model to represent both pairwise interactions and singletons in a unified framework. Additionally, the proposed approach can be extended to account for situations where a pedestrian interacts with many other pedestrians by setting the right hand side of (21) to be equal to an integer greater than one.

V. Experimental Results

The proposed MRF modeling and inference approach is demonstrated using a self-created Cornell Campus database and the PETS09-S2L1 database taken from the Multiple Object Tracking Challenge (MOTC) benchmark [19]. Both databases are comprised of several outdoor streaming videos with multiple pedestrians. For comparison purpose, this paper also implements an existing algorithm known as dominant sets (DS) proposed in [20]. As a first step, two validation scenarios are considered using the Cornell Campus database: the Campus Road scenario (Fig. 3) and Ho Plaza scenario (Fig. 4). Subsequently, the MRF approach is compared to DS by obtaining average inference accuracy results across multiple validation videos. Finally, the PETS09-S2L1 database is used to demonstrate that inferred pedestrian relationships influence pedestrian trajectories and may, thus, be able to improve upon pedestrian behavior and trajectory predictions.

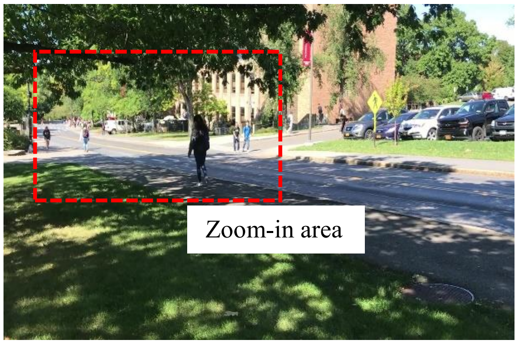


Fig. 3. Campus Road region of interest (ROI).

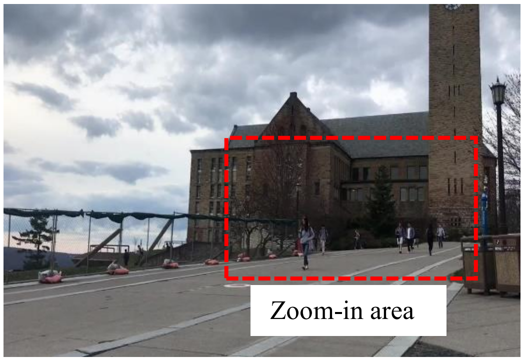


Fig. 4. Ho Plaza region of interest (ROI).

In the Campus Road region of interest (ROI), five pedestrians are detected and their bounding boxes labeled and extracted into MRF nodes, as shown in Fig. 5.(a). In this video clip, only pedestrians 4&5 are interacting socially (ground truth). After the video observation matrix is obtained from the video clip, the pedestrian trajectories are plotted as curves of the same color as the corresponding bounding boxes and MRF nodes in Fig. 5.(b)-(c), where the units of FOV relative position are in pixels. The pairwise pedestrian interactions obtained by the DS algorithm are shown as arc connections between the corresponding nodes, over time (video clip index), in Fig. 5.(b). It can be seen that, because of non-discriminating proximity features, the DS algorithm erroneously infers interactions between pedestrians 1&2. Instead, the MRF algorithm presented in this paper is able to correctly infer the interaction between pedestrians 4&5 as well as the singletons without errors (Fig. 5.(c)).

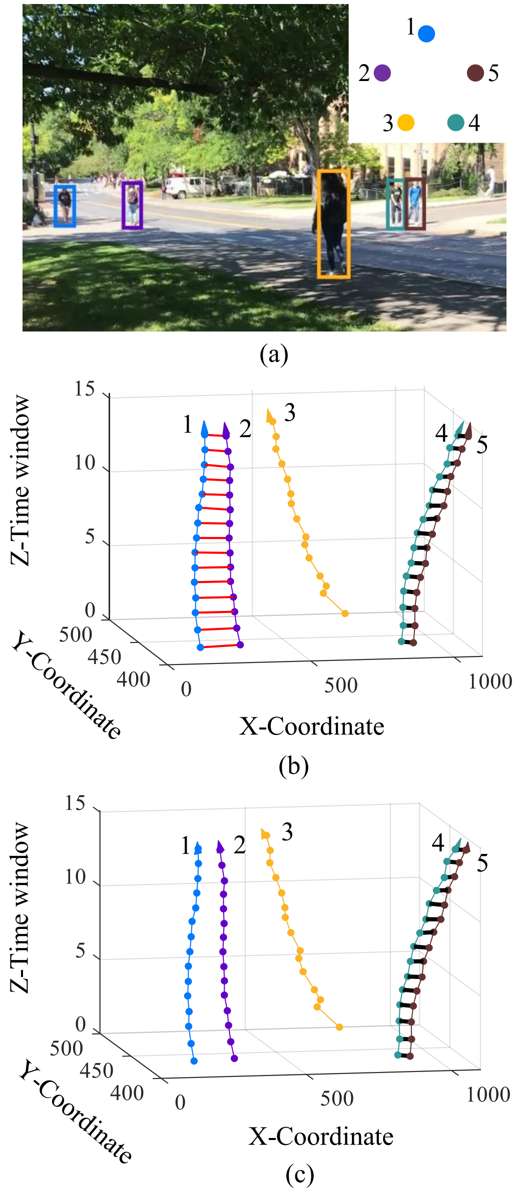
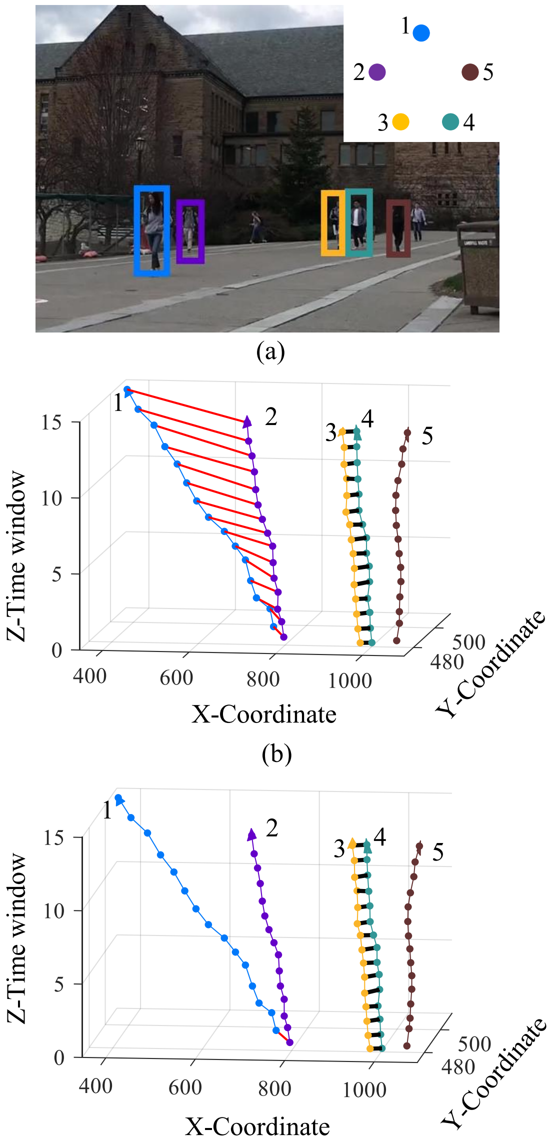


Fig. 5. Campus Road sample video frame and pedestrian (labeled) bounding boxes (a) and pedestrian trajectories and relationships inferred via DS (b) and MRF (c) algorithms.

In the Ho Plaza ROI, five pedestrians detected are detected and their bounding boxes labeled and extracted into MRF nodes, as shown in Fig. 6.(a). In this video clip, only pedestrians 3&4 interact socially (ground truth). The pedestrian trajectories and social interactions inferred by the DS and MRF algorithms are plotted in Figs. 6.(b) and 6.(c), respectively. The DS algorithm once again produces an erroneous result indicating that pedestrians 1&2 are interacting for the duration of the entire video. On the other hand, the proposed MRF approach (Fig. 6.(c)) only presents an error for the first video clip (V_1), when it incorrectly infers that pedestrian 1&2 interact. The MRF error is caused by inaccurate position information due to the relative FOV coordinates measured in pixels in the camera image plane. Hence, at V_1, when pedestrians 1&2 are very far away from the camera, they appear to walk together and interact. Immediately after V_1, the MRF algorithm corrects its MAP estimate and accurately predicts all pedestrian interactions thereafter. This type of error can be prevented either by including depth information (as provided by an RGB-D camera), or by transforming the pedestrian FOV-relative position into inertial frame using landmarks and camera information.



(a)

(b)

(c)

Fig. 6. Ho Plaza sample video frame and pedestrian (labeled) bounding boxes (a) and pedestrian trajectories and relationships inferred via DS (b) and MRF (c) algorithms.

A. Performance Comparison

The performance of DS and MRF algorithms was evaluated and compared by considering several validation videos and computing an average inference accuracy, where the inference accuracy is defined as the percentage of interactions inferred correctly in each validation video. As summarized in Table I, the average inference accuracy of the proposed MRF approach is very high at 94.6%, and it significantly outperforms the clustering-based DS algorithm proposed in [5]. While DS obtains more true positives than MRF, it also brings about many more false positives, indicating that it is far less capable of identifying lonesome pedestrians. Also, unlike the proposed MRF approach, which jointly infers and models the interaction structure across all the pedestrians, the DS algorithm only computes interactions iteratively. In particular, the DS algorithm first finds a dominant interaction and, then, it removes the dominant interacting pairs, repeating the process until all remaining pedestrians have been considered. Additionally, the proposed MRF method obtains much higher inference accuracy for short video clips (results not shown for brevity).

TABLE I
INFERENCE ACCURACY COMPARISON

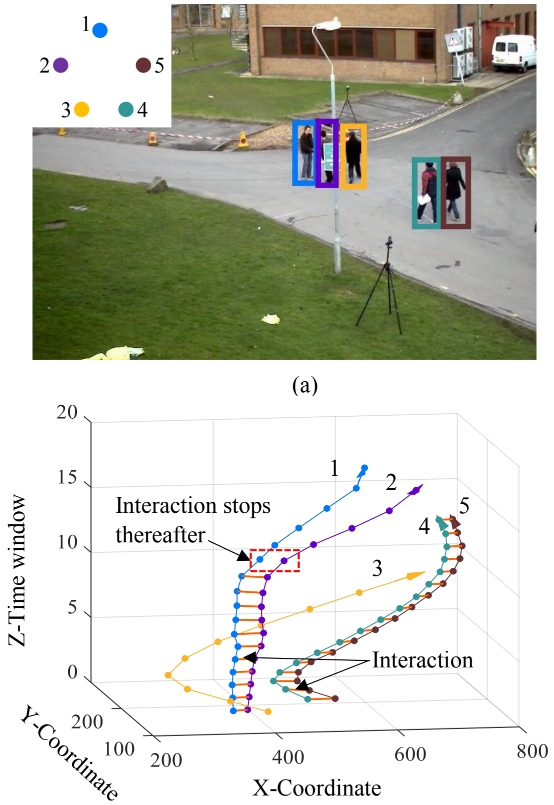
Method	True Positive	False Positive	True Negative	False Negative	Inference Accuracy
DS	186	158	66	0	61.5%
MRF	174	8	214	14	94.6%

B. Behavior and Trajectory Prediction

Ten videos drawn from the Cornell Campus database and the MOTC PETS09-S2L1 database are used to demonstrate the influence of social interactions and familiarity on the pedestrian trajectories. By measuring the relative distance variance (RDV) between every pair of trajectories it can be seen that trajectories of interacting pedestrians are highly correlated (low RDV), while those of non-interacting pedestrians are not (high RDV).

Consider, as an example, the trajectories of five pedestrians that are extracted along with bounding boxes from one MOTC video [19], shown in Fig. 7. In this video, pedestrians 1&2 interact initially, and, then, by the tenth frame, the interaction ceases and their trajectories diverge (Fig. 7). Pedestrians 4&5 interact for the entire duration of the video and, therefore, their trajectories evolve very similarly. Pedestrian 3 is lonesome and, thus, it can be seen that his/her trajectory is uncorrelated with the others.

The above qualitative observations are validated by computing the RDV of every pair of pedestrian trajectories in ten videos, $\mathcal{V}_1, \ldots, \mathcal{V}_{10}$, obtaining the results in Table II. The average RDV for non-interacting pedestrians is much higher than that of interacting pedestrians, as also indicated by the ratio of non-interacting over interacting pairs (Table II). Therefore, it can be concluded that correctly inferring



(a)

(b)

Fig. 7. MOTC sample video frame and color-coded pedestrian nodes and bounding boxes (a), and ground-truth trajectories and interactions (b).

pedestrian interactions can aid in the prediction of their future trajectories and behaviors.

VI. Conclusions

Knowledge of pedestrian social interactions and behaviors is key toward improving the efficiency and accuracy of trajectory prediction algorithms. The Markov random field approach developed in this paper utilizes spatial and temporal evidence obtained from streaming video to jointly reason and infer the social interaction structure of multiple pedestrians in a scene. New Markov random field structure and algorithms are developed in order to cast interaction inference as an energy minimization problem. With the proposed definition of energy function in terms of relative pedestrian position, speed, and orientation, energy minimization can be solved efficiently as an integer linear program. The approach is tested using the Cornell Campus database and the Multiple Object Tracking Challenge database in order to demonstrate its flexibility and robustness to different settings and environments. The experimental results show that the proposed Markov random field method significantly outperforms the dominant sets clustering algorithm, achieving an average interaction inference accuracy of 94.6%.

Acknowledgment

This work is supported by the Office of Naval Research under awards N0014-19-1-2266 and N00014-17-1-2175.

TABLE II
Average Relative Distance Variance (RDV) of Pedestrian Trajectories

Video	Interacting Pedestrians (IP)	Non-interacting Pedestrians (NP)	Ratio (NP/IP)
$\mathcal{V}_1$	10.9	1303.4	119.6
$\mathcal{V}_2$	3.6	2193.0	609.2
$\mathcal{V}_3$	3.3	19.2	5.82
$\mathcal{V}_4$	4.8	703.7	146.6
$\mathcal{V}_5$	0.8	1660.5	**2075.6**
$\mathcal{V}_6$	6.4	136.3	21.3
$\mathcal{V}_7$	4.3	951.5	221.3
$\mathcal{V}_8$	18.1	958.5	53.0
$\mathcal{V}_9$	15.5	874.4	56.4
$\mathcal{V}_{10}$	28.2	859.9	30.5

References

[1] M. Maurer, J. C. Gerdes, B. Lenz, H. Winner, *et al.*, "Autonomous driving," *Berlin, Germany: Springer Berlin Heidelberg*, vol. 10, pp. 978–3, 2016.

[2] E. Ohn-Bar and M. M. Trivedi, "Looking at humans in the age of self-driving and highly automated vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 1, pp. 90–104, 2016.

[3] H. Bai, S. Cai, N. Ye, D. Hsu, and W. S. Lee, "Intention-aware online pomdp planning for autonomous driving in a crowd," in *2015 ieee international conference on robotics and automation (icra)*. IEEE, 2015, pp. 454–460.

[4] P. Trautman and A. Krause, "Unfreezing the robot: Navigation in dense, interacting crowds," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 797–803.

[5] K. N. Tran, A. Gala, I. A. Kakadiaris, and S. K. Shah, "Activity analysis in crowded environments using social cues for group discovery and human interaction modeling," *Pattern Recognition Letters*, vol. 44, pp. 49–57, 2014.

[6] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *CVPR 2011*. IEEE, 2011, pp. 1345–1352.

[7] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 104.

[8] J. Gemerek, S. Ferrari, B. H. Wang, and M. E. Campbell, "Video-guided camera control for target tracking and following," *IFAC-PapersOnLine*, vol. 51, no. 34, pp. 176–183, 2019.

[9] H. Wei, P. Zhu, M. Liu, J. P. How, and S. Ferrari, "Automatic pan–tilt camera control for learning dirichlet process gaussian process (dpgp) mixture models of multiple moving targets," *IEEE Transactions on Automatic Control*, vol. 64, no. 1, pp. 159–173, 2018.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[11] M. Moussaıd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, "The walking behaviour of pedestrian social groups and its impact on crowd dynamics," *PloS one*, vol. 5, no. 4, 2010.

[12] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[13] S. Nowozin, C. H. Lampert, *et al.*, "Structured learning and prediction in computer vision," *Foundations and Trends® in Computer Graphics and Vision*, vol. 6, no. 3–4, pp. 185–365, 2011.

[14] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[15] J. M. Gonfaus, X. Boix, J. Van de Weijer, A. D. Bagdanov, J. Serrat, and J. Gonzalez, "Harmony potentials for joint classification and segmentation," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3280–3287.

[16] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.

[17] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.

[18] S. M. Azar, M. G. Atigh, A. Nickabadi, and A. Alahi, "Convolutional relational machine for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7892–7901.

[19] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *arXiv:1504.01942 [cs]*, Apr. 2015, arXiv: 1504.01942. [Online]. Available: http://arxiv.org/abs/1504.01942

[20] M. Pavan and M. Pelillo, "Dominant sets and pairwise clustering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 1, pp. 167–172, 2006.