Predicting Social Interactions for Visual Tracking

Xu Yan
xyan6@uh.edu
Ioannis A. Kakadiaris
ioannisk@uh.edu
Shishir K. Shah
sshah@central.uh.edu

Department of Computer Science
University of Houston
Houston, TX 77204-3010
USA

Human interaction dynamics are known to play an important role in the development of robust pedestrian trackers that are applicable to a variety of applications in video surveillance. Traditional approaches to pedestrian tracking assume that each pedestrian walks independently and the tracker predicts the location based on an underlying motion model, such as a constant velocity or autoregressive model. Recent approaches have begun to leverage interaction, especially by modeling the repulsion force, among pedestrians to improve motion predictions. However, human interaction is more complex and is influenced by both repulsion and attraction effects. This motivates the use of a more complex human interaction model for pedestrian tracking.

In this paper, we propose a novel visual tracking method by leveraging complex social interactions that decomposes social interactions into multiple potential interaction modes. It is hard to define the social interaction mode for every pedestrian pair without an explicit knowledge of their intent. To address this point, we treat pedestrian interaction as a linear combination of potential social interaction modes with dynamically adjusted weights at different moments in time. We decompose social interaction into potential social interaction modes and quantify them using a social force model [3] as shown in Eq. 1. For pedestrian i, the total force is given by:

$$\vec{F}_i = \sum_{n=1}^{q_i} w_i^{d_n} \vec{F}_i^{d_n},\ d_1, d_2, \ldots, d_{q_i} \in D_i, \tag{1}$$

where d_n denotes the n^{th} social interaction mode, $\vec{F}_i^{d_n}$ represents social forces based on mode d_n, $w_i^{d_n}$ is a weighting coefficient, D_i is the set of potential social interaction modes, and q_i is the number of social interaction modes. For any pedestrian, the decomposition is defined by first building social links with other pedestrians. The attraction and repulsion effects of the link are denoted by $\{+\}$ and $\{-\}$, respectively, and a social interaction mode is denoted by a set of social links' effects. Each pedestrian's social interaction set is expanded to have the same size as the maximum number in the social interaction set. This is done to ensure that the number of decomposed motion models remain the same across all tracked pedestrians.

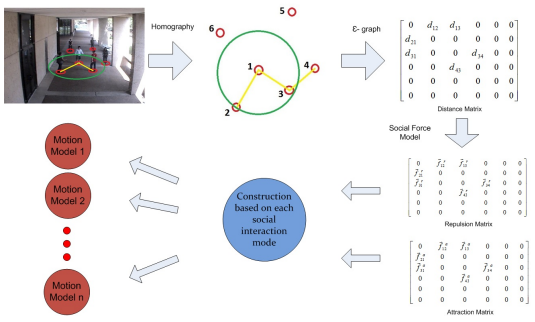

Figure 1: Depiction of the social interaction decomposition framework

Our method estimates an accurate value of posterior probability by designing a sophisticated motion model $p(x_t|x_{t-1})$ within the framework of a Bayesian tracker. Following the decomposition method by Kwon and Lee [4], the motion model is designed as the weighted linear combination of its basic components and given by:

$$p(x_t|x_{t-1}) = \sum_{n=1}^{N} w_t^n p_n(x_t|x_{t-1}),\ and\ \sum_{n=1}^{N} w_t^n = 1, \tag{2}$$

where $p_n(x_t|x_{t-1})$ denotes the n^{th} basic motion model, w_t^n is the weighting variable at time t, and N is the number of decomposed motion models, which is equal to the cardinality of the social interaction set. Let $\vec{c}_t = [\vec{c}_{1,t}, \vec{c}_{2,t}, \ldots, \vec{c}_{P,t}]^T$ and $\vec{v}_t = [\vec{v}_{1,t}, \vec{v}_{2,t}, \ldots, \vec{v}_{P,t}]^T$ denote the positions and velocities of the pedestrians, respectively. Given $\vec{F}^{d_n} = [\vec{F}_1^{d_n}, \vec{F}_2^{d_n}, \ldots, \vec{F}_P^{d_n}]^T$, the next position information for pedestrians can be predicted. Using a constant velocity motion model, the motion prediction is defined as $\vec{c}_t = \vec{c}_{t-1} + \vec{v}_{t-1}\Delta t$ in which Δt is the time interval between two frames. Incorporating the social interaction force for prediction, the state update at a fixed interval of time Δt is given as:

$$\begin{bmatrix} \vec{c}_t^{d_n} \\ \vec{v}_t^{d_n} \end{bmatrix} = \begin{bmatrix} \vec{c}_{t-1} + \vec{v}_{t-1}\Delta t + \frac{1}{2}\frac{\vec{F}^{d_n}}{m}\Delta t^2 \\ \vec{v}_{t-1} + \frac{\vec{F}^{d_n}}{m}\Delta t \end{bmatrix}, \tag{3}$$

where $\vec{c}_t^{d_n}$ and $\vec{v}_t^{d_n}$ denote the predicted positions and velocity of pedestrians under the interaction mode d_n, respectively. For a single pedestrian i, the motion prediction $\vec{c}_{i,t}$ is given by a set of locations predicted by each of its social interaction modes, $\vec{c}_{i,t} = \{\vec{c}_{i,t}^{d_1}, \vec{c}_{i,t}^{d_2}, \ldots, \vec{c}_{i,t}^{d_q}\}$. In this work, each basic motion model $p_n(x_t|x_{t-1})$ is modeled as a Gaussian distribution and is given by:

$$p_n(x_t|x_{t-1}) \propto \mathcal{N}(\vec{c}_t; \vec{c}_t^{d_n}, \sigma^2). \tag{4}$$

Given each decomposed motion model, a basic tracker is composed of a pair of observation and motion models. This generates a basic tracker n that uses the observation model $p(y_t|x_t)$ and the motion model $p_n(x_t|x_{t-1})$ describing a specific social interaction mode d_n. The total number of basic trackers for a pedestrian i is the same as the total number of social interaction modes in the set D_i. A Markov chain is constructed for each of the trackers. The state space is updated according to the MAP estimate obtained via the Metropolis Hasting algorithm [2]. The Interactive Markov Chain Monte Carlo [1] algorithm is leveraged to sample across the basic trackers and combine their sampling results. We test our method on videos from unconstrained outdoor environments and compare it against popular multi-object trackers.

Figure 2: The tracking results of our dataset over representative frames: our proposed tracker (row1), BPF (row2), MCMC (row3), and VTD (row4).

[1] J. Corander, M. Ekdahl, and T. Koski. Parallel interacting MCMC for learning of topologies of graphical models. *Data Mining and Knowledge Discovery*, 17(3):431–456, 2008.

[2] W.K. Hastings. Monte Carlo sampling method using Markov Chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[3] D. Helbing and P. Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, 1995.

[4] J. Kwon and K.M. Lee. Visual tracking decomposition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1269–1272, San Francisco, June 2010.