MODELING LOCAL BEHAVIOR FOR MULTI-PERSON TRACKING

A Dissertation

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Xu Yan

May 2014

MODELING LOCAL BEHAVIOR FOR MULTI-PERSON TRACKING

Xu Yan

APPROVED:

Shishir Shah, Chairman
Dept. of Computer Science

Edgar Gabriel
Dept. of Computer Science

Christoph Eick
Dept. of Computer Science

Ioannis Kakadiaris
Dept. of Computer Science

Saurabh Prasad
Dept. of Electrical & Computer Engineering

Dean, College of Natural Sciences and Mathematics

Acknowledgements

First, I would like to thank to my advisor, Dr. Shishir Shah, for his guidance, discussion, support, and huge help during my time at University of Houston. It has been my fortune to be his student and the best experience of my life working with him for the past five years. Second, I would like to thank my co-advisor, Dr. Ioannis Kakadiaris, who give me a lot of valuable suggestion in academic research. I would also thank my comittee memebers, whose contributions and suggestions were greatly appreciated and served to improve the quality of this dissertation. I also thank members of Quantitative Imaging Lab for all the disscussion and cooperation. I must also thank my family, my parents and parents-in-law, for their love, support, patience, and understanding. And last but not least my wife, Can Li for your accompaniment, love, and encougragement. Special thank to my coming baby, without you and your mom, this Ph.D. degree would not have been possible.

MODELING LOCAL BEHAVIOR FOR MULTI-PERSON TRACKING

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Xu Yan

May 2014

Abstract

Multiple-pedestrian tracking in unconstrained environments is an important task that has received considerable attention from the computer vision community in the past two decades. Accurate multiple-pedestrian tracking can greatly improve the performance of activity recognition and analysis of high level events through a surveillance system.

Traditional approaches to pedestrian tracking build a motion prediction model to track the target. With improvements in object detection methods, recent approaches replace the motion prediction stage and track targets by selecting among the outputs of a detector. To incorporate the merit of traditional and recent approaches, we have developed a novel approach using an ensemble framework that optimally chooses target tracking results from that of independent trackers and a detector at each time step. The compound model is designed to select the best candidate scored by a function integrating detection confidence, appearance affinity, and smoothness constraints.

To further improve the tracking performance we focus on the design of a novel motion prediction model. Human interaction behavior is known to play an important role in human motion. We present a novel tracking approach utilizing human collision avoidance behavior, which is motivated by the human vision system. The model predicts human motion based on modeling of perceived information. An attention map is designed to mimic human reasoning that integrates both spatial and temporal information. We also develop an enhanced tracker that models human group behavior using a hierarchical group structures. The groups are identified by a bottom-up social

group discovery method. The inter- and intra-group structures are modeled as a two-layer graph and tracking is posed as optimization of the integrated structure.

Finally, we propose another novel tracking method to unify multiple human behavior. To investigate the effects of potential multiple social behaviors, we present an algorithm that decomposes the combined social behaviors into multiple basic interaction modes, such as attraction, repulsion, and no interaction. We integrate these multiple social interaction modes into an interactive Markov Chain Monte Carlo tracker and demonstrate how the developed method translates into a more informed motion prediction, resulting in robust tracking performance.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Multiple-pedestrian tracking in unconstrained environments is an important task that has received considerable attention from the computer vision community in the past two decades. A number of approaches that address this problem have been proposed [75, 73] for its importance in applications related to surveillance, human activity recognition, and video retrieval. Accurate multiple-pedestrian tracking can greatly improve the performance of activity recognition and analysis of high level events through a surveillance system. However, the complexity of human motion poses several challenges to the accuracy and precision of any tracking system. In the context of video surveillance, human motion can be thought of as blob motion in which arms and legs are difficult or unnecessary to localize. At this scale, the study of human motion predominantly involves cues related to space and environment, and

we can expect to recover how people move from place to place. Accordingly, the recovery of motion pattern of people facilitates a measure of social phenomena among interacting individuals [30]. Interpersonal distance cues have their basis in the seminal findings that people tend to organize the space around them in four concentric zones associated with different degrees of intimacy [27]. The spatial organization of people within these concentric zones is dominated by relationships between interacting individuals [54]. Hence, it is the encoding of social relationships along with tracking methods that has been most commonly exploited in recent years to model human motion.

Visual tracking of multiple targets in complex scenes captured by a monocular, potentially moving, and uncalibrated camera is a very challenging problem due to measurement noise, cluttered background, uncertainty of the target motion, occlusions, and illumination changes [73]. While traditional methods for tracking have focused on improving the robustness of motion models and predictive filters, recent advances in methods for object detection [23, 68, 67] have led to the development of a number of *tracking-by-detection* [4, 29, 20, 62, 14, 9, 5, 38] approaches. These methods try to first apply a learned discriminative model to detect objects in each frame independently, and then associate detections across frames to identify each object's unique spatio-temporal trajectory. However, varying visual properties of the object of interest often results in tracking drift, imprecise detection, missing detection and occlusion as shown in Figure 1.1. Hence, the resulting association problem has to be resolved by inferring between-object interactions using incomplete data sets. Several

(a) Tracking drift (b) Imprecise detection (c) Missing detection (d) Occlusion

Figure 1.1: Examples of outputs from a tracker (in red box) and a detector (in green box).

approaches have been proposed to address this problem by optimizing detection assignments in the spatio-temporal context [76, 34, 5, 12], while other methods have focused on achieving the necessary precision by coupling a robust tracker that can update its predictive model (motion and object attributes) guided by the detection confidence and discriminative features obtained from multiple cues [14, 20, 74]. The improved tracking performance reported by these methods indicates that such a combination is desirable. Nonetheless, the positive or negative contribution of the chosen predictive model and the detector at each time step within the combination term is not well understood. Further, it is not guaranteed that each term will have an equivalent contribution towards tracking a target and the weighting parameters chosen empirically could deteriorate the tracking result in previously unobserved scenarios.

1.2 Challenges

1.2.1 Ensemble of Detection and Tracking

Traditional trackers that depend on the appearance model and motion prediction perform poorly in the presence of abrupt motion changes and cause template drifts.

The true target gradually shifts away from the tracking template because of the error in the motion model. In addition, tracking drifts and photometric variations make it hard to maintain unique identities among targets and cause frequent identity switches. On the other hand, detection results suffer from long-term occlusion, dynamic backgrounds and low-resolution images. Since outputs from either the detector or the tracker can be sparse and unreliable, one solution to alleviate the problem is to create an abundant number of potential candidates to increase the probability of finding a more accurate candidate for the target of interest. For example, visual tracker sampler [40] samples a large number of trackers from the tracker space dynamically to compensate for target variations. Other approaches have also tried to combine the tracker and detector together [14, 74] but limited the role of the detector so as to assist the tracker as a confidence measurement tool. The benefit of using outputs from the tracker and detector directly as association candidates for the tracked target, however, has never been fully exploited.

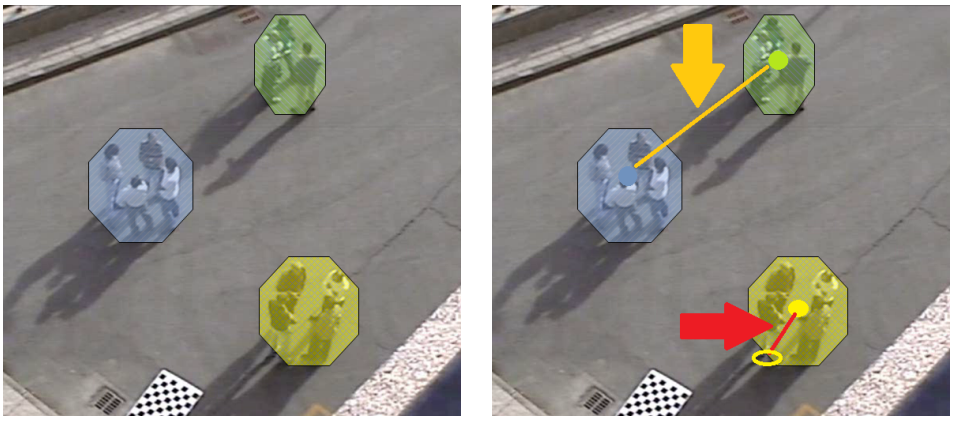
1.2.2 Visual Perception in Multi-person Tracking

Unlike the traditional motion model, the social behavior model, in essence, treats human motion as the result of both a person's intention and their interaction with environment rather than the outcome of a motion dynamics model alone. This is a critical aspect of tracking humans and enables incorporation of the basic understanding that human beings invariably will make motion decision based on their intent and understanding of the environment. Typical social behavior models are built on constraints over spatial proximity and treat nearby subjects and objects with equal

importance. However, a person does not plan his/her movements based on a holistic understanding of the scene but reasons about it based on the local field of visual perception [31]. Therefore, we propose building a perception-based motion model from the first-person perspective. Intuitively, a person does not react to all subjects in his/her perspective with equal intensity. For example, a person will react strongly to a person moving faster in their direction as compared to someone moving slower. In other words, a person moving quickly towards one will take priority in one's perception and hence in their motion planning. We argue that people's attention has two kinds of variations: (1) spatial variations that are related to subjects that are near or far; and (2) temporal variations that are related to subjects that are moving fast or slow. An attentive vision based motion model is more realistic and beneficial for improving multi-person tracking.

1.2.3 Group Structure in Multi-person Tracking

Among various measures of social interaction, social grouping provides an indication of how humans engage and orient themselves to exhibit a group activity. With regards to motion behaviors, a social group can be inferred from pedestrian trajectories. In the case of multiple people in a scene, social groups can be indicated through relationships within a group and relationships across groups. We can denote the structure across groups as an intra-group structure and the structure within a group as an inter-group structure as shown in Fig. 1.2. Social grouping can guide tracking by assuming that humans in a group will maintain their spatial structure in the coming moments. The key benefit of taking advantage of social grouping is two



(a) Social Group (b) Group Structure

Figure 1.2: Examples of social group and group structures. The red arrow shows an inter-group structure and the yellow arrow shows an intra-group structure.

fold, one is to handle human occlusions within a group and second is to minimize the search space for data association. Modeling both inter-group and intra-group structure can fully exploit these key benefit.

1.2.4 Social Interaction in Multi-person Tracking

The integration of social relationships to address the dynamics of human motion has its origin in the social force model [33] that applies a fluid flow analogy to the dynamics of pedestrians. It is primarily a physical model that captures a continuous phenomena where humans are considered to react to energy potentials caused by other pedestrians and static obstacles, while trying to keep a desired speed and motion direction. Recently proposed local motion models such as linear trajectory avoidance model (LTA) [49] or human motion prediction model [43] demonstrate that leveraging social relationships can improve tracking performance. The effect of the

social relationships can take two forms: 1) attraction effects, and 2) repulsion effects. The attraction and repulsion effect can be characterized as the tendency to move toward or away from objects. Repulsion effect has been leveraged in most existing tracking methods, but modeling of both effects of social relationships simultaneously remains challenging. Modeling motion based on repulsion effects alone excludes the possibility of people's intent to meet and only captures the intent of avoiding collisions. Nevertheless, unconstrained environments would typically involve people with motion dynamics explained under both repulsion and attraction effects.

1.3 Research Goals

First, we argue that results obtained from the tracker and detector generate redundant association candidates and can complement each other in different scenarios. For example, a drifting tracking result can be corrected by the detection result (Figure 1.1(a)) and an imprecise detection result can be replaced by a better tracking prediction (Figure 1.1(b)). In the case of missed detection (Figure 1.1(c)) and occlusion (Figure 1.1(d)), the prediction power of the tracker may help to maintain the position and identity of the tracked target. Similarity measurement of appearance model alone is unreliable and the exploration of the interplay among multiple cues in a tracking environment yields promising results [14, 20]. So we propose a ensemble framework to optimize the association by selecting outputs of a detector or a tracker at each time step to increase the overall tracking accuracy and precision. Instead of using detection and classification results to guide the tracker, we treat

the tracker and object detector as two independent identities and we keep both their results as association candidates for the tracked target. In each frame, we select the best candidate and assign it to the tracked target. The assignment is scored by a function integrating detection confidence, appearance affinity, and smoothness constraints imposed using geometry and motion information. The approach exploits the discriminative power of the tracker and detector. The weighting factor of each term used in the score function is discriminatively trained. The method of data mining for hard negative examples [23] is applied to handle a very large set of artificially generated negative samples.

Secondly, we define human motion as a direct consequence of human attentive vision system. The problem is then transformed into a human attentive vision modeling problem, which operates in a virtual simulation world that has the same physical world coordinates as the real world. We simulate the virtual vision and get the first-person perspective image. Such transformation facilitates intuitive analysis of human perception and reaction to subjects in the environment and induces a more realistic motion model. Then we propose an attentive vision model that approximates the spatial and temporal variance of human attention. The combined attention map enables motion path prediction of a person without explicit knowledge of other person's motion. Our method identifies regions of high interest from subject's attention map that guides the estimation of subject's next movement and serves as a novel feature in a person tracking framework. The attention feature is integrated into a tracking-by-detection framework and improves the tracking performance.

Furthermore, most tracking systems with social grouping have overlooked the

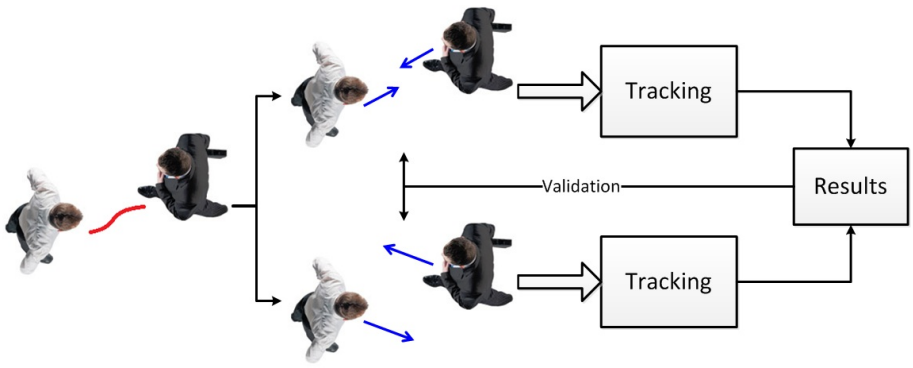


Figure 1.3: Depiction of the process flow.

structure across groups while the individual-group relation or the inter-group structure has been exploited. Current systems account for the inter-group structure as a important constraint to refine the target search during tracking. However, intra-group structure may change in a dramatically different manner from inter-group structure and could provide rich information to further improve tracking. For instance, the structure for individuals in the same group may stay unchanged or change minimally while exhibiting slow movements or in stationary states. On the other hand, the structure between groups may change significantly if groups are moving in different directions. To fully leverage the social grouping context, inter-group and intra-group structures should be modeled and updated in a more integrated approach. We present an approach to do so by extending the model of Zhang *et al.* [77] and show that it improves multiple person tracking.

Last but not least, the intent of pedestrians produces different social relationships in which the intent of avoidance is explained by the repulsion effect and the intent of approach is explained by the attraction effect. The intent varies over time, thus

motion prediction of corresponding trackers should be adjusted dynamically depending on the current interaction environment. One main limitation of current work is that the motion dynamics of a target is modeled using a fixed motion model, typically a first-order approximation. Thus, it fails to model the complex motion that is affected by elaborate pedestrians' intent and corresponding interactions. Our approach focuses on how to incorporate the temporally varying pedestrian interaction into a dynamic motion model without explicit knowledge of local social relationships. Although the content of interaction is unknown, the intent of pedestrians can be assumed to belong to a finite set which combines the intent of avoidance and approach. The finite set of intent generates a finite set of interactions. We propose to decompose complex pedestrian interaction into a finite set of interactions, where the decomposition is motivated by the work of Kwon and Lee [39]. With the decomposition of complex pedestrian interactions, we present a visual tracker based on a pedestrian dynamic model that combines both the intent driven terms of avoidance and approach. The main idea of our method is illustrated in Fig. 1.3. Local interaction modeling guides the tracking result. Conversely, the tracking output validates the content of social interactions. Consider a simple scene consisting of two pedestrians. We model the local interaction between them under the intent of either avoidance or approach. Based on the modeling, our approach predicts two possible motions for each pedestrian. Then it searches the best tracking result by sampling pedestrians' state space. On the other hand, the best tracking result validates the intent under which local interaction effects contribute more accurately to prediction using a linear search strategy.

1.4 Dissertation outline

The organization of the remainder of this dissertation is as follows. We begin in Chapter 2 by presenting a literature review of existing approaches to multi-person tracking. In Chapter 3, we present the ensemble framework for optimal selection. Chapter 4 proposes a visual perception modeling method for multi-person tracking. Hierarchical group structures model us explained in Chapter 5. Chapter 6 describes the social interaction decomposition strategy and the compound social interaction based tracker. Finally, the last chapter summarizes the dissertation highlights and its contributions with a discussion on future work.

Chapter 2

Related Work

2.1 Multi-person Tracking

Multi-person tracking is a fundamental problem for many computer vision tasks, such as video surveillance and activity recognition. Various approaches have been proposed to address this problem [71]. Previous tracking algorithms mainly exploit two aspects including coping with targets' appearance variance and modeling complex targets' motion. To account for appearance variation of the target caused by change of illumination, deformation and pose, a large amount of work has been proposed [79, 2, 28, 55, 45, 8] and these methods perform well and get good results. However, the dynamics of target and interaction between targets is much less explored. The state space of targets is affected by motion of target and interaction of targets. While a dynamic model is used mainly to reduce the search space of state space, it affects the tracking results especially when multiple targets undergo complex interacting

motion. The early tracking algorithms adopt constant velocity model or acceleration model, such as [50] which leverages second-order auto-regressive dynamics. Most recent tracking algorithms incorporate random walk model [55, 45] which models the targets' dynamics as Brownian motion. Recently, various approaches that account for interaction among people to improve visual tracking have been developed. Motion behaviors are distinctly encoded dependent on the tracking environment, primarily differentiated in terms of densely crowded scenes and more informal interacting environments. Social behavior in densely crowded scenes is dominated by the overall motion of all individuals in the scene while informal interacting environments allow for more locally complex dynamics. In the following, we mainly focus on recent work in tracking by detection and sampling based trackers.

2.1.1 Tracking by Detection

Building on the success of state-of-the-art object detection methods, object tracking appears to be "easier" to achieve if best matching detection targets can be transitively linked. However, due to the numerous false positives and missed targets in detection results, local data association based on affinity measures between contiguous detections is hard to achieve, hence limiting the ability to find a unique trajectory for each tracked target without drifting [69, 29]. On the other hand, global data association tries to solve the problem by optimizing the linkage problem of multiple trajectories simultaneously [76, 34, 5]. Brendel *et al.* [15] used maximum weighted independent set to converge to a data association optimum. Andriyenko *et al.* [6] combined discrete with continuous optimization to solve both data association and

trajectory estimation. Butt *et al.* [18] use Lagrangian relaxation to transfer the global data association to solvable min-cost flow problem. Since global methods tend to be computationally expensive, they usually start by detecting short tracklets and iteratively associating them into longer tracks. Leibe *et al.* [41] propose to couple the output of detector and trajectory estimation, but trajectories ultimately rely on detections. To overcome the difficulties faced by the global association approaches, Breitenstein *et al.* [14] proposed to deal with the detection uncertainty in a particle filtering framework in which unreliable detection information is complemented by the prediction power of the tracker. In order to increase the association confidence, a boosted classifier is trained online to assess the similarity of tracker-detection pairs. Independent from the detector's output, the classifier term improves the robustness of the tracking result. This coupling framework has also been applied to challenging sports videos [74], which uses a vote-based confidence map to localize people, and the motion model is estimated by the optical flow.

2.1.2 Sampling-based Trackers

The interaction among multiple targets contributes to the complexity of state space. Particle filters based tracking methods perform well in handling non-Gaussianity and multi-modality of the distribution of targets' state [35]. Previous works extended the particle filters framework for multiple targets by either leveraging multiple independent trackers [57] or increasing the joint state space to include multiple targets [53]. The first approach is computationally tractable but can not handle interacting targets. In the second approach, the complexity of computation expands exponentially

with the number of targets since the joint state space becomes increasing large. To overcome these problems, Khan *et al.* [36] and Zhao *et al.* [78] proposed the Markov chain Monte Carlo (MCMC) based particle filter tracker which reduces the computational cost in a high dimensional state space and make sampling-based tracker feasible for multi-target tracking. Kwon *et al.* [39] utilize the interactive MCMC (IMCMC) to sample multiple basic tracker space, which requires a relatively small number of samples by exchanging information between chains and is capable of solving combinatorial problem. This dissertation also leverages the IMCMC based particle filter tracking framework to search the best targets' states. In the real-world environment, a fixed number of trackers are insufficient to cope with complicated tracking environment. Thus, our approach changes the total number of basic trackers and does the sampling in a changing joint target space efficiently. To the best of our knowledge, a similar algorithm is proposed in [40]. It samples the pre-defined basic tracker pool and maintains the number of tracker as small as possible while our tracker changes the number of basic trackers based on the social interaction mode at each time instant.

2.2 Visual-Attention Modeling

By mimicking the human vision system, computational visual-attention modeling is investigated by psychologists, microbiologists, and computer scientists. A number of computational models of attention are proposed and can be categorized based on

whether they are biological, purely computational, or hybrid [24]. All plausible biological methods are directly or indirectly inspired by cognitive concepts. In contrast, Ma *et al.* [44] proposed a method based on local contrast for generating saliency maps that is not based on any biological model. Achanta *et al.* [1] had incorporated both biological and computational parts in their method. Our work falls in the area of purely computational methods. All the aforementioned tracking works treat the social behavior from surveillance camera view angle instead of understanding social behavior from subject's own viewpoint. In this dissertation, we model the target motion behavior from the first-person view and utilize it for multiple target tracking. To the best of our knowledge, no previous tracking method has leveraged first-person perspective.

2.3 Local Behavior Modeling

Khan *et al.* [36] proposed to model the level of interaction between targets at any time step according to the percentage of pixel overlap between the bounding boxes of the corresponding targets. The joint behavior of interacting targets is defined according to the interaction potentials measured by a graph-based Markov Random Field (MRF). Specifically, the interaction potential is measured in the log domain, expressed by means of the Gibbs distribution. The discrete-choice model proposed by Antonini *et al.* [7] was targeted for human tracking and assumes that an individual pedestrian makes a choice from a discrete set of velocity options at each time step. The options are defined by four elements: a choice set, a set of attributes describing

the alternatives, a set of socio-economic characteristics describing the decision maker and a random term ϵ to capture the correlation structure between alternatives. This model treats other pedestrians as occupants of a potential path. The position and the direction of walk contributes to the decision making within the choice set. Building on the social force model [33], the linear trajectory avoidance (LTA) model [49] accounts for local repulsion effects among pedestrians to estimate the energy potential and predict motion paths. The LTA model incorporates the human dynamics attributed to the behavior where a pedestrian tries to avoid running into other pedestrians while approaching ones destination. The underlying assumption of LTA model is that each pedestrian has their own global destination and all the other pedestrians are treated as moving obstacles. The motion of a pedestrian is predicted by minimization of an energy functional that accounts for the pedestrians' position, speed and angle. Luber *et al.* [43] extended the use of the local repulsion effect to include physical and social constraints of the environment. Specifically, the model combined the personal intention forces and the repulsion social forces of other pedestrians and physical obstacles to define the total force induced over each pedestrian. The individual pedestrian motion was then predicted based on the sum of forces inferred through a closed-form solution. Choi *et al.* [20] considered both local repulsion effects and group motion dynamics within a joint prediction model. The group motion dynamics capture the effect that some people walking close to each other tend to keep walking together. It models the repulsion effects based on the distance between two targets in the 3D space. The group motion is modeled based on the motion correlation within a group while comparing both the similarity of velocities and the distance among

each other. Qin *et al.* [52] and Bazzani *et al.* [10] exploited the social group effect associated with the tracking performance.

2.4 Open Challenges

Several open challenges for multi-person tracking can be draw from related work:

- The redundancy and diversity between tracking and detection is not leveraged for robust multi-target tracking. Previous approaches perform tracking by associating the output of either the detector or basic tracker only.
- To the best of our knowledge, visual perception has never been exploited for multi-person tracking which can help understand human motion in a realistic way.
- While the individual-group relation or the inter-group structure has been exploited, modeling both inter-group and intra-group structures has never been done.
- None of the previous models incorporate local-attraction effects which is one main cause of abrupt motion change into local motion prediction. Broadly speaking, all models based on the use of local repulsion effects only [49, 43, 20] are applicable in environments where pedestrians are moving in a crowd and do not deviate from their global destination. In contrast, the modeling of more complex local interactions among people exhibited through both local repulsion and attraction effects has not been adequately explored.

Chapter 3

An Ensemble Framework for Optimal Selection

3.1 System Overview

Our system is initialized with a human detector and several independent human trackers. Each independent tracker deals with one target. As illustrated in Figure 3.1, after collecting redundant candidates from outputs of both the detector and independent trackers in testing stage, the hierarchical data association step tries to optimize the association between tracked targets and candidates. We reduce the association problem to an assignment problem. To manage time complexity, we adopt a greedy-search based association framework using the score matrix between candidates and targets as detailed in Section 3.2. The score is computed by the dot product between a set of learned weights and features extracted from multiple

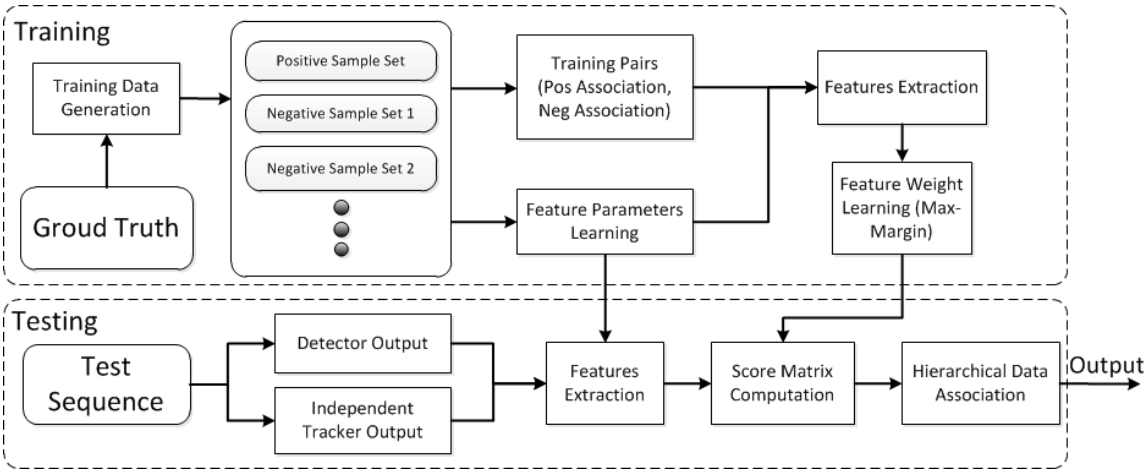


Figure 3.1: Framework of our tracking system.

cues. In addition to color histogram, optical flow, and motion features, we learn an additional target classifier to measure the object detection confidence. Those weights are trained with a max-margin framework, which is designed to give high affinity score for associating candidates with true tracked targets and low score when tracked targets are associated with drifting, false positive candidates or candidates belonging to different targets. To learn the weight parameter, positive samples are obtained from the ground truth and a large number of negative samples are artificially generated to prevent sample selection bias as described in Section 3.3. We test our method using publicly available datasets under different challenging conditions and demonstrate superior tracking results that outperform the state-of-the-art algorithms, particularly in terms of accuracy.

Figure 3.2: Using outputs from both the tracker and the detector, our algorithm selects the best candidate and associates it to the tracked target. Candidates from the tracker are shown in red boxes and candidates from the detector are in green boxes. Solid boxes represent the best candidate selected by the algorithm.

3.2 Ensemble Model

We formulate the multi-object tracking problem as a sum assignment problem that associates tracking candidates obtained from outputs of the tracker and detector to tracked targets of interest. Let $\mathbb{S} = \{s_1, \ldots, s_m\}$ be all tracked targets, where m is the number of objects currently being tracked. Let $\mathbb{TR} = \{tr_1, \ldots, tr_m\}$ indicate the set of all independent trackers. In this dissertation, a color based particle filter is implemented for each independent tracker $tr_i \in \mathbb{TR}$. Each particle filter tracker deals with one target and $\mathbb{T} = \{t_1, \ldots, t_m\}$ represents the output of particle filter trackers. The detector's output is denoted as $\mathbb{D} = \{d_1, \ldots, d_n\}$, where n is the number of detection outputs. $\mathbb{D} \cup \mathbb{T}$ represents all $m + n$ tracking candidates. The aim of the system is to find the optimal assignment for all tracked targets $\mathbb{S}$ in each frame t, which is measured by the association score of the form:

$$\begin{aligned} &\arg\max_{\{j\}} \sum_{i=1}^{m} \beta \cdot \Phi_t(x_i^j) \\ &i \in \mathbb{S}, j \in \mathbb{D} \cup \mathbb{T} \\ &s.t.\ \forall x_a^p, x_b^q \quad p \neq q \ if \ a \neq b, \end{aligned} \tag{3.1}$$

where x_i^j indicates that the candidate j is assigned to tracked target i, $\{j\}$ denotes a set of selected candidates, β is a vector of model parameters which is learned as presented in Section 3.3, $\Phi_t(\cdot)$ represents the association feature set in the current frame, and $\beta \cdot \Phi_t$ is the score function. The proposed formulation tries to find optimal links between tracked targets and candidates provided by both the tracker and the detector by assigning at most one candidate to at most one target, and the assignment is evaluated by the affinity score defined in Equation 3.1. The association feature set

$$\Phi(x_i^j) = \left[\phi_1(j), \phi_2(x_i^j), \phi_3(x_i^j), \phi_4(x_i^j), \phi_5(x_i^j), \phi_6(x_i^j)\right]$$

combines information from different feature spaces, namely the classification confidence, the color histogram, the motion feature and the optical flow feature. Each component is described below in detail.

Classification confidence. The classification confidence (ϕ_1) is proportional to the likelihood that the object of interest appears at a given position, and the confidence is derived from the classification result of a binary classifier introduced to gain additional robustness to our discriminative framework [14]. The classifier scores a feature vector x with a dot product function $\omega \cdot x$, where ω is a vector of weighting parameters and x is the feature vector extracted from a given image patch. w is trained with a max-margin framework, the details of which are provided in Section 3.3. The feature vector x is the concatenation of multi-scale HOG [22] and LBP [3] feature sets. The HOG feature is extracted from a two-level image pyramid. We perform PCA on each HOG feature to manage the curse of dimensionality and improve prediction performance. The cumulative energy threshold for selecting

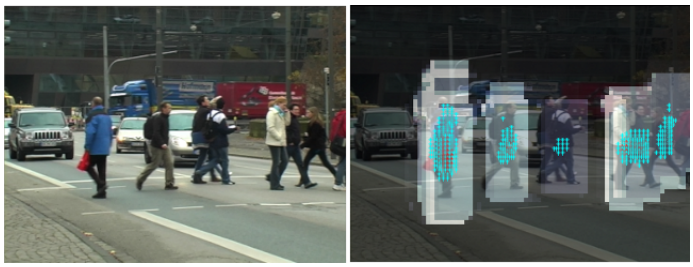


Figure 3.3: Visualization of the classification confidence map.

eigenvectors is set at 95%. For image patches with negative score, their classification confidence is set to zero. For those with positive score, the score is normalized by its ranking among scores of positive training samples. For example, an image patch with classification confidence of 0.95 will have higher score than 95% of positive training samples. Figure 3.3 illustrates the confidence map after applying the binary classifier for human detection. As can be seen, areas of detection targets yield high confidence value.

Color histogram. The 3D color histogram is built in the Red-Green-Intensity (RGI) space with 5 bins per channel. Given the training pairs, we perform a kernel density estimate for the target and candidate. The similarity between two kernels $g(x_c)$ and $g(x_t)$ is measured by the Bhattacharyya coefficient B, and the likelihood is given by:

$$\phi_2 \propto \exp(-\lambda(1 - B[g(x_c), g(x_t)])) \ , \tag{3.2}$$

where λ is set to be 5.

Motion feature. The speed, object scale, and angle represent the motion feature of objects. The speed is modeled by the Normal distribution; $\phi_3 \propto f_s(\frac{s_t}{s_{t-1}}; \mu^s, \sigma^s)$, where $\frac{s_t}{s_{t-1}}$ is the speed ratio between two frames and f_s is the probability density

function. The angle likelihood is modeled by the *von Mises* distribution [61], which is formulated as:

$$\phi_4 = \frac{e^{\kappa^a cos(\theta - \mu^a)}}{2\pi I_0(\kappa^a)} \ , \tag{3.3}$$

where $I_0(.)$ is the modified Bessel function of order zero. The scale likelihood is modeled by the Normal distribution; $\phi_5 \propto f_l(l; \mu^l, \sigma^l)$, where l is the scale between two frames and f_l is the probability density function. Model parameters $\{\mu^s, \sigma^s, \mu^a, \kappa^a, \mu^l, \sigma^l\}$ are learned from positive training samples.

Optical flow feature. The optical flow is precalculated according to [16]. The dominant motion of each region is encoded by a 2D histogram that quantifies both the magnitude and angle of the motion with 10 bins and 8 bins, respectively. The Bhattacharyya coefficient B is used to measure the similarity of histograms $\{H_t, H_c\}$ between the target and candidate. The optical flow score function is given by $\phi_6 \propto \exp(-\tau(1 - B[H_t, H_c]))$, where τ is set to be 5.

3.2.1 Hierarchical Data Association

Our algorithm employs a hierarchical association strategy to solve Equation 3.1 by progressively associating outputs of independent trackers and the detector to tracked targets. We use the word "active" to distinguish a target without being occluded. The association hierarchy consists of three levels. In each level, the assignment is obtained by the Hungarian algorithm. At the first level, it finds the best association between active targets and all candidates. At the second level, we try to associate occluded targets to all unassigned candidates of the detector. If the best association

score is below a threshold, we move the occluded target to the third level, in which, it will be linked to one of the unassigned candidates of the indpendent trackers based on the association score. The details of the algorithm are described in Algorithm 1.

Greedy search by Hungarian Algorithm. Given m targets, we solve the assignment problem for n detector's candidates and m independent trackers' candidates through the Hungarian algorithm. The score matrix is defined as:

$$S = \underbrace{\left[\begin{array}{cccc|cccc} s_1^1 & s_1^2 & \dots & s_1^n & s_1^{n+1} & -\infty & \dots & -\infty \\ s_2^1 & s_2^2 & \dots & s_2^n & -\infty & s_2^{n+2} & \dots & -\infty \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s_m^1 & s_m^2 & \dots & s_m^n & -\infty & -\infty & \dots & s_m^{n+m} \end{array}\right]}_{\text{Detector's Candidates}} \underbrace{\qquad\qquad}_{\text{Trackers' Candidates}}$$

The score in S is computed by $s_i^j = \beta \cdot \Phi(x_i^j)$ where i and j are row and column, respectively. The negative infinity value of the off-diagonal components represents self-association rule of the tracking result [34] as it is designed to be linked to one specific target only.

Occlusion Handling. We set the "enter" and "exit" regions along image borders after the first two frames in a typical surveillance setting similar to [14]. If the best association score for a target is below a threshold in the first level association, it will be marked as "occluded". In addition, an assigned target in the first level with a lower classification confidence score than a threshold will also be marked as "occluded". We activate an occluded target only if its associated candidate returns a classification confidence score greater than a threshold. For an occluded target, the association feature set is not updated until it becomes active again. An occluded

target will be deleted if it stays in the "exit" region for more than 5 frames or remains "occluded" for more than 20 frames in the "non-exit" regions. Deletion of an active target depends on its position with respect to the "exit" region.

Algorithm 1: Association Framework

Input: Targets $\{1, \ldots, i, \ldots, m\}$, Candidates $\{1, \ldots, j, \ldots, (n+m)\}$.
Output: Association Results.

Compute the association feature for all active targets and candidates;
Compute the score $s_i^j = \beta \cdot \Phi(x_i^j)$ and assign it to the score matrix;
Apply the Hungarian algorithm to solve the assignment problem;
for *each assignment* (i, j) **do**
    if $s_i^j <$*threshold or* $\phi_1(j) <$*threshold* **then**
        invalidate the assignment
    end
end
Recompute the score matrix for all occluded targets and unassigned detection candidates;
Apply the Hungarian algorithm to solve the assignment problem;
for *each assignment* (i, j) **do**
    if $s_i^j <$*threshold or* $\phi_1(j) <$*threshold* **then**
        invalidate the assignment
    end
end
if *active target is assigned* **then**
    update the feature set;
else
    active target is set as occluded;
end
if *occluded target is assigned* **then**
    occluded target is set as active and its feature set is updated;
end
For all unassigned occluded targets, associate them to corresponding tracking results

3.2.2 Detector and Independent Trackers

We use a state-of-the-art deformable part based detector to detect the occurrences of human targets in every frame [23]. The tracker we use is similar to [50] where the particle's observation model is built upon the RGI color histogram [14]. We keep track of two frames for updating the observation template: one is the first frame that the target appears in and the other is the latest frame in which the target is in the "active" state. A new independent tracker is initialized for a target that has higher classification confidence value than the threshold in two continuous frames and has no existing tracker associated to it. The observation template is updated based on the first frame object appear and the recent frame object appear. Second order autoregressive dynamics is leveraged to propagate the particles:

$$x_{t+1} = Ax_t + Bx_{t-1} + Cv_t, v_t \propto \mathcal{N}(0, \Sigma) \tag{3.4}$$

Where matrices A, B, C and Σ defining the motion dynamics. All the parameter is experimental results. A, B, C keep the same for all the dataset. Σ is increased for sport video set. A new tracker is initialized for an object that has two subsequent detections in overlapping region, which are not associated to an exiting tracker. The enter region for the first two frame is the whole image. For all the other frames, new tracker is only initialized in enter region which is shown in Figure 3.4. The initial samples are drawn based on a Normal distribution centered at the detection box. A tracker is terminated when the object is set as deleted.

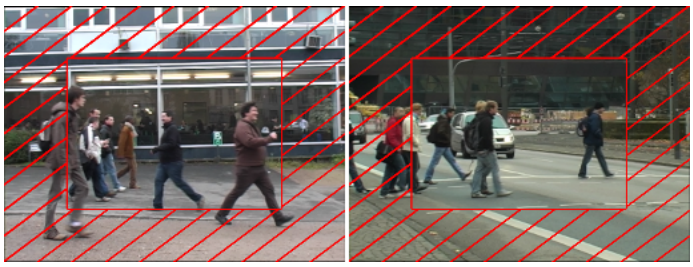


Figure 3.4: Examples of the enter and exit regions after the first two frames.

3.3 Discriminative Learning

The model parameters β in Equation 3.1 is learned discriminatively. Consider a training set $\{(y_i, \Phi(o_i))\}_{i=1}^{N}$, where $y_i \in \{+1, -1\}$ is the label and $\Phi(o_i) \in \mathbb{R}^n$ is the feature vector extracted from the training instance o_i that includes an assigned candidate. We are trying to learn a model that assigns the score to the instance with a function of the form $\beta \cdot \Phi(\cdot)$, where β is a vector of model parameters. The formulation, in analogy to classical SVMs, leads to the following optimization problem:

$$f(\beta) = \min_{\beta} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{N} \ell(\beta; (\Phi(o_i), y_i)) \ , \tag{3.5}$$

where $\ell(\beta; (\Phi(o_i), y_i)) = \max\{0, 1 - y_i \langle \beta, \Phi(o_i) \rangle\}$ is the hinge loss function and the constant C is chosen experimentally as the weight for the penalty. Stochastic subgradient method [23, 60] is applied for solving this problem.

Typically positive samples are given and we manually generate the "hard" negative samples. For training, negative samples are randomly generated for three different kinds of scenarios: tracking drift, false positive and mismatch. As shown in

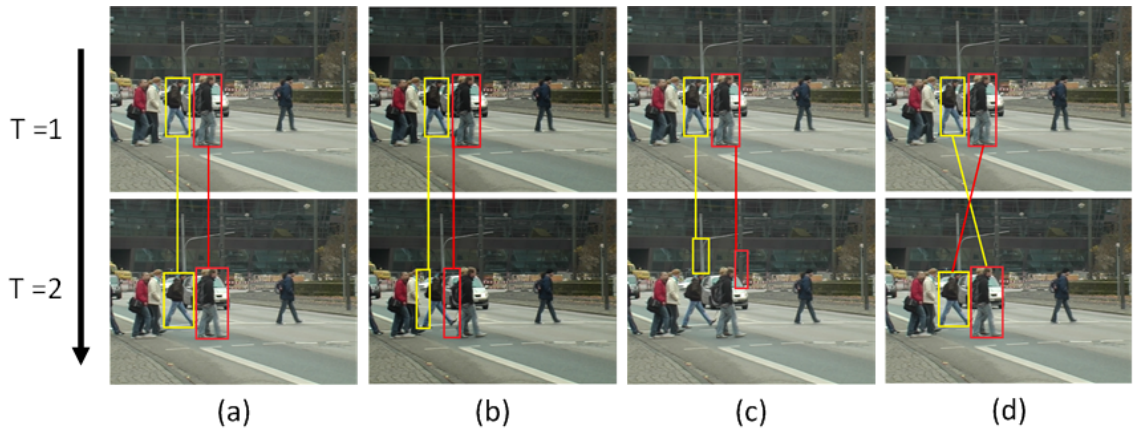


Figure 3.5: Positive and negative samples generated for the tracking problem: (a) correct tracking, (b) tracking drift, (c) false positive, (d) mismatch.

Figure 3.5(b), samples for tracking drift are picked as image patches that have between $0\% \sim 25\%$ overlaps with the ground truth. False positive accounts for cases where there is no overlap between tracked objects in the ground truth and candidates obtained from the tracker and detector (Figure 3.5(c)). Mismatch represents identity switch, in which a tracked target is connected to a wrong candidate (Figure 3.5(d)). Compared with the number of positive samples, the number of negative instances is very large. To deal with a large set of samples, we apply the data-mining algorithm proposed in [23] for training our model efficiently.

3.4 Experiments

3.4.1 Datasets

We evaluate our tracking algorithm on four public challenging datasets: TUD Crossing, TUD Campus, ETHZ Central and UBC Hockey [14]. The video in the UBC Hockey dataset is taken by a moving camera and static cameras are used for videos in other datasets. These four datasets presents a wide range of challenges due to heavy inter-person occlusion, poor image quality, and low image contrast between targets and the background. Videos in these datasets also cover different viewpoints and capture various types of movements. In all experiments, we define "entry" and "exit" zone manually for each sequence and no other scene or calibration knowledge is leveraged. We employ the discriminatively trained deformable parts model [23] as the human detector. The detector uses publicly available and pre-trained model for TUD Crossing and TUD Campus datasets. The deformable parts model is re-trained for ETHZ Central and UBC Hockey datasets to boost the detection rate as the quality of images is much poorer in these videos. None of the video frames in the dectector training are used for testing.

3.4.2 Parameter Training

We obtained the model parameter β as described in Section 3.3. We use the same parameter for TUD Crossing and TUD Campus datasets, which is trained on the first 25 frames of the TUD Crossing video. β is trained for ETHZ Central by using

the first video sequence in the dataset and our tracking algorithm is tested on the second video sequence. For UBC Hockey, the first 25 frames are used for training. None of the video frames in the parameter training set are used for testing.

3.4.3 Quantitative Evaluation

We adopt CLEAR MOT metrics [13] to evaluate the tracking performance of our algorithm. Key measurements of the metrics include: precision score (MOTP) measured by intersection over union of bounding boxes, and an accuracy score (MOTA) which is composed of false negative rate (FN), false positive rate (FP), and number of ID switches (ID Sw.). Results of our algorithm are reported in Table 3.1 (shown in top row) after conducting experiments on aforementioned four datasets. In general, the result shows that our approach achieves high tracking accuracy with very few number of ID switches. In our experiments, false positives usually are caused by the drift of occluded targets since it is hard to update the motion model in time during occlusion. For example, the rapid change of movements in the UBC Hockey video increases the chance of false positives. The failure of the human detector is the main reason for false negatives in the result since several persons are not detected and corresponding independent trackers are not initialized in the video. A typical detection failure happens when occlusions persist over several video frames. For example, as shown in Figure 3.6, one of the persons sitting in the lower-right corner is never detected. Occlusion is also the culprit for ID switches. If a newly detected person bears similar appearance with an occluded target, the ID of the occluded one may be mistakenly assigned to another target.

Dataset	MOTP	MOTA	FP	FN	ID Sw.
TUD Crossing	70.77%	**89.38%**	**1.09%**	**9.33%**	**2**
TUD Crossing[14]	**71.0**%	84.3%	1.4%	14.1%	**2**
TUD Campus	**67.76%**	**84.82%**	**0.0%**	**15.18%**	**0**
TUD Campus[14]	67.0%	73.3%	0.1%	26.4%	2
ETHZ Central	**71.49%**	**75.4%**	0.36%	**24.24%**	**0**
ETHZ Central[14]	70.0%	72.9%	**0.3%**	26.8%	**0**
ETHZ Central[41]	66.0%	33.8%	14.7%	51.3%	5
UBC Hockey	**71.61%**	**91.75%**	1.76%	**6.49%**	**0**
UBC Hockey[14]	57.0%	76.5%	1.2%	22.3%	**0**
UBC Hockey[47]	51.0%	67.8%	**0.0%**	31.3%	11

Table 3.1: CLEAR MOT evaluation results on four datasets. Our results are in the top row for each dataset. The best results are in bold.

For comparison, we list the results of three competing approaches for these sequences: (i) On-line Multi-Person Tracking-by-Detection [14] on TUD Crossing, TUD Campus, ETHZ Central and UBC Hockey; (ii) Coupled detection and trajectory estimation [41] on ETHZ Central; (iii) Boosted particle filter [47] on UBC Hockey. As shown in Table 3.1, we outperform the competing approaches on all datasets in terms of tracking accuracy. As for the tracking precision, our results are comparable with the best reported performance measures.

To fully evaluate the benefit of the ensemble tracking-by-detection framework proposed in this dissertation, we also present the performance of component-wise analysis. The default method used the output of both the part-based detector and independent particle filter trackers to accomplish data association. Variant (a) leverages output of particle filter trackers alone as tracking candidates while variant (b) leverages output of only the part-based detector. As shown in Table 3.2, the default method performs better in term of accuracy, false positive, false negative and ID

Dataset	MOTP	MOTA	FP	FN	ID Sw.
TUD Crossing (**Default**)	**70.77%**	**89.38%**	**1.09%**	**9.33%**	**2**
TUD Crossing (a)	63.06%	58.13%	20.04%	21.63%	19
TUD Crossing (b)	69.46%	82.14%	6.85%	10.81%	36
TUD Campus (**Default**)	67.76%	**84.82%**	**0.0%**	**15.18%**	**0**
TUD Campus (a)	62.80%	47.19%	18.15%	33.99%	**0**
TUD Campus (b)	**70.08%**	59.74%	17.16%	22.44%	3
ETHZ Central (**Default**)	71.49%	**75.4%**	**0.36%**	**24.24%**	**0**
ETHZ Central (a)	59.25%	26.74%	37.97%	34.94%	23
ETHZ Central (b)	**75.59%**	72.91%	1.96%	24.78%	7
UBC Hockey (**Default**)	71.61%	**91.75%**	**1.76%**	**6.49%**	**0**
UBC Hockey (a)	58.32%	80.41%	4.85%	14.56%	26
UBC Hockey (b)	**73.42%**	82.84%	3.41%	13.57%	10

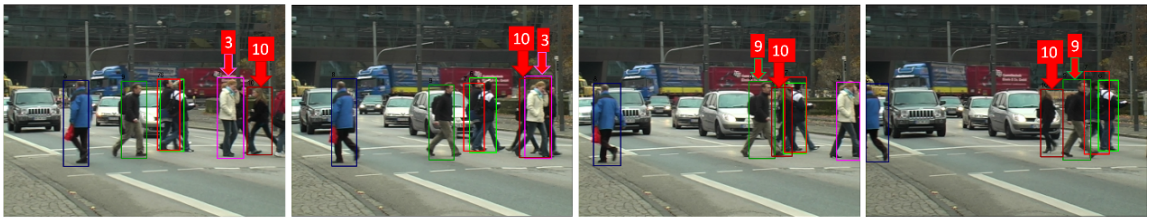
Table 3.2: CLEAR MOT evaluation results on component-wise evaluation of our approach. Variant (a) leverages output of the independent trackers only. Variant (b) leverages output the detector only. The best results are in bold.

switches over the result of variants due to the optimal selection of output of both components. The lower precision score of the default method in three of four datasets is related to the way MOTP is computed. Since MOTP only measures the positional deviation of detected targets from their ground truth, an increase in the number of detected targets can lead to lower overall precision. This is the case since the default method is able to track a greater number of targets than either of the variants.

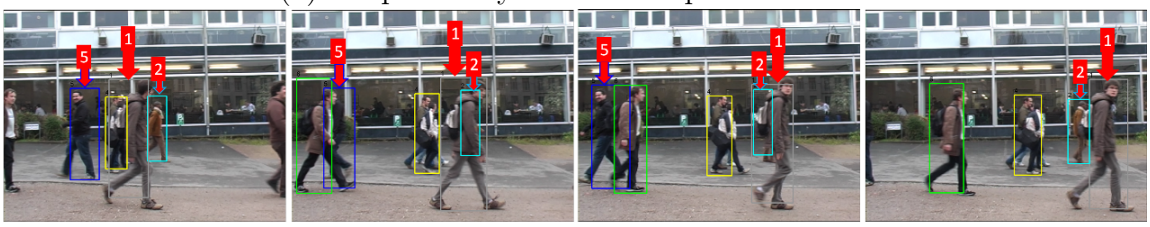
3.4.4 Qualitative Evaluation

Figure 3.6 shows our qualitative evaluation. The first row presents the ability of our approach to keep the identity for target #10 even when the target has been occluded by multiple targets in the sequence. The scenario in the second row shows

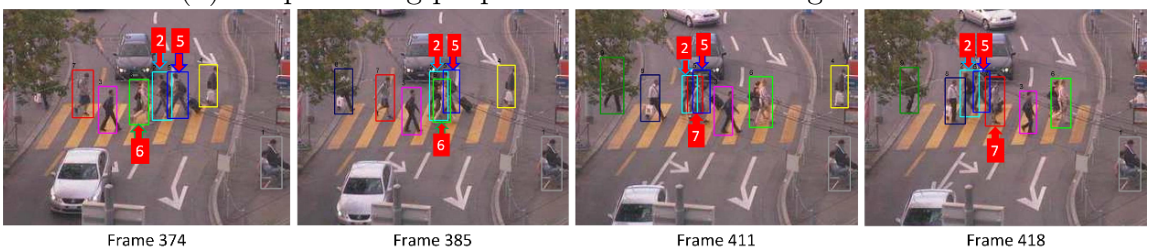
that we can keep updating the location of occluded targets by using the tracker's prediction in the case of missing detections. The third row demonstrates that our tracker can differentiate targets well when they are very close to each other. The last row shows a sequence shot from a moving camera. Although the motion model loses its accuracy due to abrupt changes of movements, our tracker can correct tracking drifts by switching to associate detection results to tracked targets.



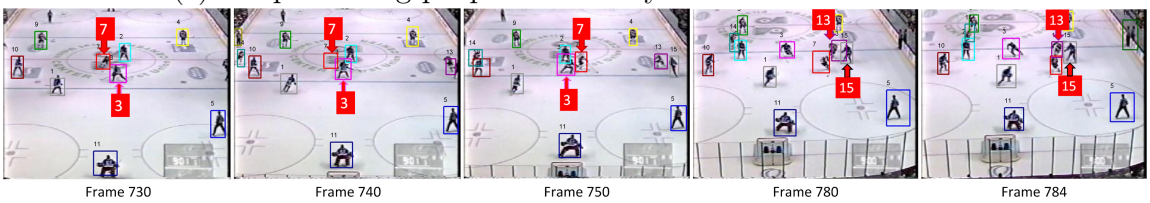
(a) Keep identity under multiple occlusions



(b) Keep tracking people in the case of missing detections



(c) Keep tracking people when they are close to each other



(d) Correct the tracking drift in the moving camera scenario

Figure 3.6: Tracking results of our approach on TUD Crossing, TUD Campus, ETHZ Central and UBC Hockey datasets.

Chapter 4

Modeling Human Visual Perception for Multi-person Tracking

Benfold and Reid [11] utilized a person's head pose to locate areas of attention to guide surveillance systems. However, this information was not incorporated into a multi-person tracking framework. In our case, to visualize the scene from each person's point of view we utilize the ***virtual vision simulation*** [63] so that the scene can be rendered graphically and further used to simulate a first-person view assuming the camera to be located at the head height for each person in the scene. Figure 4.1(a) shows the real world, (b) shows the virtual scene, and (c) shows the first-person view of person in the red bounding box in (a) and (b). Finally, Figure 4.1(d) shows the retinal mapping of the first-person view of the specific person based

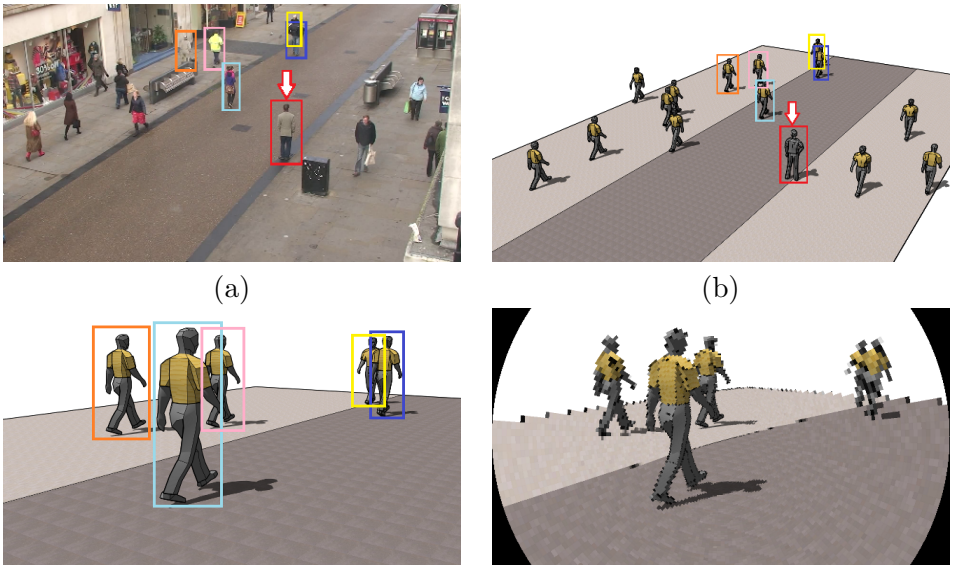


Figure 4.1: Examples of reconstructed virtual world. (a) The original surveillance image, (b) the virtual vision image, (c) the first-person view image from the person under the arrow, (d) the retinal mapping image. The bounding boxes in (a,b,c) with same color represent the same person.

on the log-polar transformation wherein the center of the first-person view image is assumed to be the focal point.

We generate "attention maps" of the simulated first-person view that guides the person's motion as shown in Figure 4.2. The static attention map is built based on human detection, which treats human subjects in the first-person perspective as obstacles. The dynamic attention map is derived from optical flow displacement of human subjects in a person's view. Human subjects further away or moving slowly exhibit smaller optical flow displacements than those in closer proximity or moving fast. Besides, the optical flow displacement from first-person perspective implicitly incorporates the effect of motion direction in which humans subjects moving towards

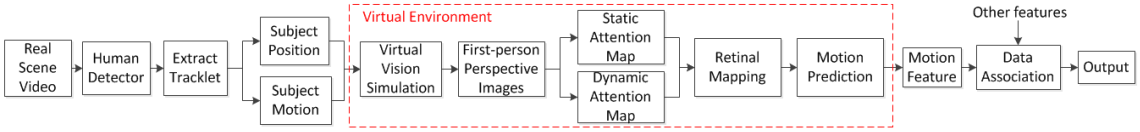


Figure 4.2: System Framework. The components in red outline are implemented in virtual environment.

a person exhibit bigger optical flow than those moving away in the same speed from the same position due to fact the person approaching show bigger image pixel shift than going away. After combining static and dynamic attention maps, retinal mapping is overlaid on the combined map to mimic human retinal vision mechanism, i.e., spatial regions far from individual's visual center will have low attention and hence lower spatial resolution and vice versa for closer regions. The final attention map combines spatial and temporal variations of the scene as per the person's visual priority.

4.1 Attentive Vision Modeling

Given a configuration $C^t = \{c_i^t\}$ of subjects $(i = 1 \dots N)$ at time t, each subject is modeled as $c_i^t = (p_i^t, s_i^t, a_i^t)$, where p_i^t denotes the world coordinate position, s_i^t its speed, and a_i^t its motion angle. Our method models the human perception of each subject i at the time step t based on the configuration C^t. For simplicity, we will explain one subject's attentive vision model in a scene with a fixed number of subjects. This can easily be generalized to an arbitrary number of subjects. Unlike previous approaches, we don't assume each person's prior knowledge about other subjects' position.

4.1.1 Virtual Vision Simulation

We assume a person's consistent moving direction in the next step $t+1$ is same as the person's current motion direction. Based on the calibration of the real scene (Figure 4.1(a)) and the output of human detection, we can get the position parameter p_i^t for each subject i. The motion parameters s^i and a^i estimation will be explained in section 4.2. Here we assume we have the parameters c_i^t for each subject. To simplify the configuration, we also set every person's height as 1.7 meter and the eye position is 1.6 meter from the ground, which is also set as the first-person view camera's position. Using the configuration C^t, we construct the virtual scene as shown in Figure 4.1(b) with virtual vision simulator [63]. In the virtual scene, we are observing the human motion by putting our camera in the observer's positions. We position the camera at the person's head position and the focal length is fixed for each person, so the camera views the scene in the direction of the person's head pose. Here we assume the head pose is same as the subject motion direction. An example of a camera's view of the scene from first-person perspective is shown in Figure 4.1(c). We refer to this as the first-person perspective image. The first-person perspective image shares the same world coordinate with virtual vision image and real world image. In the following sections, all computations of attentive vision are performed on first-person perspective images. The corresponding retinal mapping image is shown in Figure 4.1(d) further explained in the following section.

(a) (b) (c)

Figure 4.3: The static map in first-person perspective view. (a) The over-head view image.(b) The first-person view image. (c) The static map is generated based on the human detection.

4.1.2 Attention map

Visual saliency is one of the most popular computational model for visual attention [37]. Similar to saliency based attention model [48], we compute an attention map that leverages both static and dynamic components of attention. The attention map is built as shown in Figure 4.2 (red outline). The first step is to construct static and dynamic maps, then to overlap retinal mapping on the combined map.

Static map. With virtual scene, all the pedestrian's motion are simulated with virtual agents that have the same velocity as the real world scene. The images of first-person perspective are collected from virtual vision simulator for frame $\{1, \ldots, i, \ldots, K\}$. Background subtraction is performed to detect the human subjects within the controlled foreground-background contrast in virtual scene [51]. The static map is built based on human detection results in frame 1. The output of human detection of frame 1 is denoted as $R^1 = \{r_1^1, \ldots, r_n^1\}$ where r_n^1 is represented by binary foreground mask. The static map of human attention is modeled as $S_s = r_1^1 \cup \ldots \cup r_n^1$.

Dynamic map. Human perception is sensitive to moving subjects and human

attentive vision treats moving subjects with different velocities differently. The dynamic map is built to address the temporal variance component of attentive vision. Optical flow (O_x^i, O_y^i) is calculated for frame $\{2, \ldots, i, \ldots, K\}$, which implicitly models the relative motion between observer's and all the other subjects' motion [16]. With the virtual vision images, the human in $\{2, \ldots, i, \ldots, K\}$ frames is detected by background subtraction and the locations are denoted as $R^{2,\ldots,i,\ldots,K}$. We set $K = 25$ in this dissertation. The motion saliency in frame i is defined as

$$M^i(x,y) = \begin{cases} sqrt((O_x^i)^2 + (O_y^i)^2) & (x,y) \in R^i \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

The final dynamic map denoted as

$$S_d(x,y) = max\{M^2(x,y), \ldots, M^i(x,y), \ldots, M^K(x,y)\}$$

combines all the motion saliency, which is determined by taking the maximum of motion intensity. A dynamic map example for one person is shown in Figure 4.4(a).

Static-dynamic map combination. We hypothesize that the human perception drives attention to specific areas when the motion intensity in that region is above a certain threshold. Thus the combination of static and dynamic map is fulfilled in a motion-conditioned strategy. The combined attention map is computed as follows:

$$S(x,y) = \begin{cases} 1 & \text{if } S_d(x,y) \geq \epsilon \text{ or } S_s(x,y) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

where, ϵ denotes the threshold on motion intensity and is set to 0.1 in this dissertation. After combination, a binary mask is generated and is overlaid on the original

(a) (b) (c)

Figure 4.4: (a) The dynamic map is generated based on virtual vision simulation. (b) The combined static-dynamic attention map. (c) The combined map mask is applied on first-person perspective image.

image as shown in Figure 4.4(b) and 4.4(c). A crucial point to note here is that even though subjects receive higher perceptual attention, the regions they occupy may have lower probability as potential future target positions.

Retinal mapping. Attentive vision refers to the reaction of people according to the visual stimuli in a dynamically changing environment, which is characterized by selective sensing in space and time as well as selective processing with respect to a specific task [58]. Selection in space involves the splitting of the visual field in a high resolution area, the fovea, and a space-variant resolution area, the periphery, which are denoted as retinal mapping. Log-polar transformation is the most common method to represent visual information with a space-variant resolution [65] and to achieve retinal mapping. The log-polar transformation conserves high resolution in the center of the image and the resolution gradually decreases away from center.

We denote (x, y) for the image coordinates and $(r_{(x,y)}, \theta_{(x,y)})$ for the corresponding polar coordinates and r_{max} denotes the maximum value of $r_{(x,y)}$. The polar mapping

of image pixel (x, y) with origin (x_0, y_0) is defined as

$$r_{(x,y)} = \sqrt{(x - x_0)^2 + (y - y_0)^2}, \text{ and } \theta_{(x,y)} = tan^{-1}(\frac{y - y_0}{x - x_0}). \tag{4.3}$$

The foveal region is defined as a round disk with the radius r_0 and origin (x_0, y_0). The image in the foveal region retains uniform resolution while the non-foveal region exhibits decreasing resolution, which is also used to indicate the importance of observations . We apply the log-polar transformation on the non-foveal part of a first-person perspective image, which is defined as the ring-shaped area $r_{max} > r_{(x,y)} > r_0$. The unified retina mapping is defined as:

$$r'_{(x,y)} = \begin{cases} r_{(x,y)} & r_{(x,y)} < r_0 \\ log(r_{(x,y)}) & r_{max} > r_{(x,y)} > r_0 \end{cases} \tag{4.4}$$

and $\theta'(x, y) = \theta(x, y)$. With the transformed log-polar coordinates, the quantization is applied along θ' and r' axes that results in G and R elements, respectively. As shown in Figure 4.5, each pixel (x, y) undergoes a transform to the log-polar space and the log-polar space is quantized. The retinal mapping of combined static-dynamic attention map is computed based on the remapping of log-polar space that transforms the log-polar image back to the Cartesian space. The remapping follows the Eq. 4.3 and 4.4 utilizing the inverse mapping of θ' and r' to x' and y', respectively. Certain number of pixels will be allocated as the same intensity value due to the quantization in log-polar space. After doing so, we get the retinal mapping on combined attention map as shown in Figure 4.5(b). Another attention search map is generated for motion prediction as shown in Figure 4.5(c). For attention search map, we compute the mean of the mapped pixel locations and assign the intensity value from the log-polar space

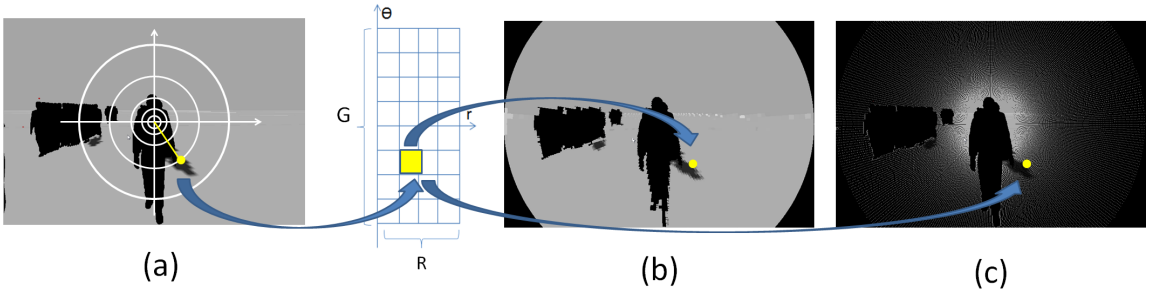


Figure 4.5: The diagram of retinal mapping. (a)The first-perspective image overlaid by static-dynamic map. (b) Retina mapping image. (c) Attention search map.

to the pixel position nearest to the computed mean position. The remaining pixels are assigned a value of zero. This allows us to generate a sparse map where the pixels that do not have a value of zero represent positions that can be probable locations for a target's next position.

4.1.3 Motion Prediction based on Attentive Vision

This dissertation assumes that people follow their intuition, which means that people will find the most feasible and most attentive point as their destination. We divide this process into two step. The first step is to find the most attentive sub-path based on attention search map (Figure 4.6(a)). A sub-path is defined as a line between two consecutive corners in the center-surround path as shown with red color in Figure 4.6(a). The probability of each sub-path in attention map is denoted as

$$P_{path} = \frac{m_{valid}}{m_{total}} \tag{4.5}$$

where, m_{valid} is the number of pixels that are not equal to zero in the attention map along the sub-path and m_{total} is the total number of pixels in the sub-path. Following

the center-surround search path, the probability map of attentive vision is generated as shown in Figure 4.6(b). The sub-paths with maximum probability are selected as most attentive sub-path by exhaustive search.

For the second step, we calculate the corresponding world coordinate of each pixel in previous optimal sub-paths. With known observer's position, the point with the shortest distance to the observer is selected as potential destination from the optimal sub-path as shown in Figure 4.7(a). The predicted human motion direction π^{att} is calculated correspondingly based on the vector from the current position to found destination and is depicted in Figure 4.7(b). This is used to guide tracking later due to the shared world coordinate between the observer and the surveillance camera's view.

4.2 Tracking Framework

To reduce the computation load and for more accurate subject motion estimation, we leverage a two-stage tracking framework. In first stage, we extracts basic tracklets $\{T_1, \ldots, T_i, \ldots, T_N\}$ for each subject i in which $T_i = \{c_i^{t_i^b}, \ldots, c_i^{t_i^e}\}$ and t_i^b and t_i^e denote the begin and end time frame of T_i. The motion parameters s_i^t and a_i^t are estimated from basic tracklets. With these parameters, we simulate the virtual vision as shown in section 4.1 and get the motion prediction with attentive vision modeling. In second stage, we combine the predicted motion feature and other features and accomplish the tracklets association.

In first stage, we leverage common method to extract basic tracklet based on

position, size and color histogram similarity in consecutive frames [52]. The color similarity constraint is also applied between current frame and first frame of tracklet. The detail of second stage is further explained in section 4.2 and 4.3.

4.2.1 Tracklet Association Formulation

We transform the tracklet association as 2D linear assignment problem on a bipartite graph. Given a set of tracklets $\mathbb{T} = \{T_1, T_2, \ldots, T_N\}$, we define a pairwise cost matrix H, in which h_{ij} denotes the cost that tracklet j is linked as first tracklet after tracklet i. The data association is formulated as

$$\arg\min_{\{i,j\}} \sum_{i=1}^{N} \sum_{j=1}^{N} h_{ij} x_{ij} \quad s.t. \begin{cases} \sum_{j=1}^{N} x_{i,j} = 1; \\ \sum_{i=1}^{N} x_{i,j} = 1; \\ x_{ij} \in \{0, 1\} \end{cases} \tag{4.6}$$

where $x_{ij} = 1$ iff tracklet j immediately follows tracklet i, otherwise, $x_{ij} = 0$. The cost is defined as the combination of five features including our attentive vision feature:

$$h_{ij} = \beta \cdot \Phi(T_i, T_j) \cdot Z(\Delta t) \tag{4.7}$$

where, $\beta = [\beta_1; \beta_2; \beta_3; \beta_4]$ is a vector of model parameters and set empirically in this dissertation, $\Phi(\cdot) = [\phi_1(\cdot), \phi_2(\cdot), \phi_3(\cdot), \phi_4(\cdot)]$ represents the association feature set, and $Z(\cdot)$ is the time gap component defined by an exponential model:

$$Z(\Delta t) = \begin{cases} \lambda^{\Delta t - 1} & 1 \leq \Delta t \leq \xi \\ \infty & \Delta t < 1 \text{ or } \Delta t > \xi \end{cases} \tag{4.8}$$

where ξ is the threshold of time gap and $\Delta t = t_j^b - t_i^e$.

4.2.2 Features Extraction

Given each tracklet pair (T_i, T_j), four features are calculated to get the association cost. The color feature ϕ_1 is build based on the 3D color histogram in the Red-Green-Intensity (RGI) space with 8 bins per channel. We perform a kernel density estimate for both the tracklets across their live frames. The similarity between two kernels $g(T_i)$ and $g(T_j)$ is measured by the Bhattacharyya coefficient B given by:

$$\phi_1 \propto \exp(-B[g(T_i), g(T_j)]). \tag{4.9}$$

The speed feature ϕ_2 is modeled by the Normal distribution: $\phi_2 \propto \mathcal{N}(\mu_j^s; \mu_i^s, \sigma_i^s)$ where, $\mu_j^s = mean(\sum_{t=t_j^b}^{t_j^e} s_j^t)$ is the average speed of T_j in its living period and μ_i^s, σ_i^s is the mean and variance of T_i's speed.

The angular likelihood is divided to two angular regions. The first one incorporates the attentive vision feature that assumes the next tracklet should appear at the predicted angle. It is modeled by the *von Mises* distribution [61], which is formulated as:

$$\phi_3 = \frac{e^{\kappa cos(\pi - \pi^{att})}}{2\pi I_0(\kappa)} \ , \tag{4.10}$$

where $I_0(.)$ is the modified Bessel function of order zero, and π denotes the motion angle between the spatial location of the middle point of tracklet i and the corresponding location of T_j. The π^{att} is our attentive vision model's predicted angle. κ corresponds to variance in a normal distribution and is set empirically. To get the informative attentive vision feature, the human motion direction history should be estimated accurately. Due to the uncertainty of detection output, we design a threshold strategy to estimate the human motion direction. When the basic tracklet

is shorter than 10 frames, we compute the average optical flow to estimate the motion direction and we rule out the region overlapped by other tracklets. Otherwise, the motion direction is computed based on tracklet position information.

The second angular feature models smooth motion and penalizes motion change. This is described by the normal distribution; $\phi_4 \propto \mathcal{N}(\mu_j^a; \mu_i^a, \sigma_i^a)$, where μ_j^a is the moving angle mean of T_j, μ_i^a and σ_i^a are the moving angle mean and variance of T_i.

4.2.3 Data association

Given the cost matrix H, we solve the assignment problem through a strategy similar to the cut-while-linking strategy proposed in [52]. The cost matrix H is extended to H^{new} to solve the initialization and termination of tracks, which is defined as,

$$H^{new} = \left[\begin{array}{cccc|cccc} h_{11} & h_{12} & \dots & h_{1N} & \tau & \infty & \dots & \infty \\ h_{21} & h_{22} & \dots & h_{2N} & \infty & \tau & \dots & \infty \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ h_{n1} & h_{n2} & \dots & h_{NN} & \infty & \infty & \dots & \tau \\ \hline \infty & \infty & \dots & \infty & \infty & \infty & \dots & \infty \\ \infty & \infty & \dots & \infty & \infty & \infty & \dots & \infty \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \infty & \infty & \dots & \infty & \infty & \infty & \dots & \infty \end{array}\right]. \tag{4.11}$$

The thresholds τ decides when a trajectory ends and is fixed for each scene. When h_{ij} exceeds τ, the link between two tracklets is cut and the track will be linked to extended columns which indicates the track terminates. The initialization of tracks is solved along with determined termination. The extended version of data association

formulation is defined as

$$\arg\min_{\{i,j\}} \sum_{i=1}^{2N}\sum_{j=1}^{2N} h_{ij}^{new} x_{ij} \quad s.t. \begin{cases} \sum_{j=1}^{2N} x_{i,j} = 1; \\ \sum_{i=1}^{2N} x_{i,j} = 1; \\ x_{ij} \in \{0,1\} \end{cases} \tag{4.12}$$

The optimal association is solved by Munkres' assignment algorithm [17].

4.3 Experiments

We evaluate how attentive vision helps to improve multi-person tracking on two public datasets: TUD stadtmitte [5] and TownCentre [12]. We follow the popular evaluation metrics [42], which includes mostly tracked trajectories (MT), mostly lost trajectories (ML), fragments (Frag) and ID switches (IDS). The TUD statmitte dataset has a short video, but with very low camera angle and frequent full occlusions among pedestrians. The TownCentre video is a high definition video with 1920×1280 resolution. This sequence is very crowded with frequent occlusion and interaction among pedestrians. The pedestrians appearing briefly at the image boundaries are excluded. We also collected a video in an outdoor uncontrolled environment. It is a high definition video with 1280×720 resolution and 1200 frames in total. This sequence is crowded with 40 trajectories in total. The activity inside is challenging for tracking algorithms since a large amount of interactions are observed among the people. Walking, skateboarding and biking activity also exists in the scene. We have manually annotated the video to identify the locations and provide unique IDs for all the people in the video.

Dataset	With attentive vision	MT	ML	Frag	IDS
TUD stadtmitte	No	60.0%	**0.0**%	3	2
TUD stadtmitte	Yes	**70.0**%	**0.0**%	**2**	**1**
TownCentre	No	81.3%	6.2%	**33**	45
TownCentre	Yes	**85.6**%	**4.8**%	43	**19**
OURS	No	47.5%	20.0%	22	21
OURS	Yes	**77.5**%	**10.0**%	**13**	**18**

Table 4.1: Component-wise evaluation on each dataset. The best result is in bold.

4.3.1 Component-wise Evaluation

To understand the benefit of the attentive vision feature proposed in this dissertation, we first present the component-wise evaluation. The baseline method turns off the attentive vision feature and re-tuning to the best performance while the default methods keep all the merits of the proposed method. Table 4.1 presents results of quantitative comparison. The default method out-performs the baseline method in most measures across all the datasets.

4.3.2 Comparative Evaluation

To compare fairly with different tracking method, we use the same detector's output. For TUD stadtmitte, we use the same detection and groundtruth provided by [72] and show comparable performance. The quantitative results are show in Table 4.2. We can see that our result is comparable or better than state-of-the-art methods. Our result is better than *Energy Min* [5], *Disc-Continue* [6] and *PRIMPT* [38] as our attentive vision incorporated model gives more informed prediction. Our approach

Dataset	Method	MT	ML	Frag	IDS
TUD stadtmitte	Andriyenko *et al.*[5]	60.0%	**0.0**%	4	7
TUD stadtmitte	Kuo *et al.*[38]	60.0%	10.0%	**0**	1
TUD stadtmitte	Andriyenko *et al.*[6]	60.0%	**0.0**%	1	4
TUD stadtmitte	Yang *et al.*[72]	**70.0**%	**0.0**%	1	**0**
TUD stadtmitte	Proposed method	**70.0**%	**0.0**%	2	1
TownCentre	Qin *et al.*[52] (a)	76.8%	7.7%	37	60
TownCentre	Qin *et al.*[52] (b)	83.2%	5.9%	**28**	39
TownCentre	Proposed method	**85.6**%	**4.8**%	43	**19**
OURS	Qin *et al.*[52]	45.0%	22.5%	24	22
OURS	Pellegrini *et al.*[49]	62.5%	15.0%	19	**16**
OURS	Proposed method	**77.5**%	**10.0**%	**13**	18

Table 4.2: Comparison of results on TUD statmitte, TownCentre and Our dataset. The best result is in bold. [52] (a) and [52] (b) represent the baseline method and proposed method in [52] respectively.

does not provide an obvious advantage over *Online CRF* [72] since this video has low camera angle and several very short tracklets, which makes it difficult to estimate the tracklet motion direction. In this case, the power of online learned appearance model in *Online CRF* gives more benefit than motion prediction. Some sample tracking results are shown in Figure 4.8.

For TownCentre dataset, we use the original detection and groundtruth provided by [12], which are used in [52], and we show improvement by incorporating the attentive vision features. The quantitative comparison is shown in Table 4.2. The results show that the attentive vision based tracking model outperforms *Basic affinity model* [52] and *SGB model* [52] in terms of MT, ML, and IDS. Fragment of trajectories under our model increased due to threshold setting in cut-while-linking strategy. Example qualitative result is shown in Figure 4.9, Figure 4.10, and Figure 4.11.

We compare our method's performance with *SGB model*. We also replace the attentive vision model in our framework with *LTA model* [49] and keep all the other components fixed. The quantitative results are shown in Table 4.2, which show that attentive vision model outperforms *SGB model* and *LTA model* in terms of MT, ML and Frag. LTA does a little better in IDS than our model. The qualitative evaluation is shown in Figure 4.12.

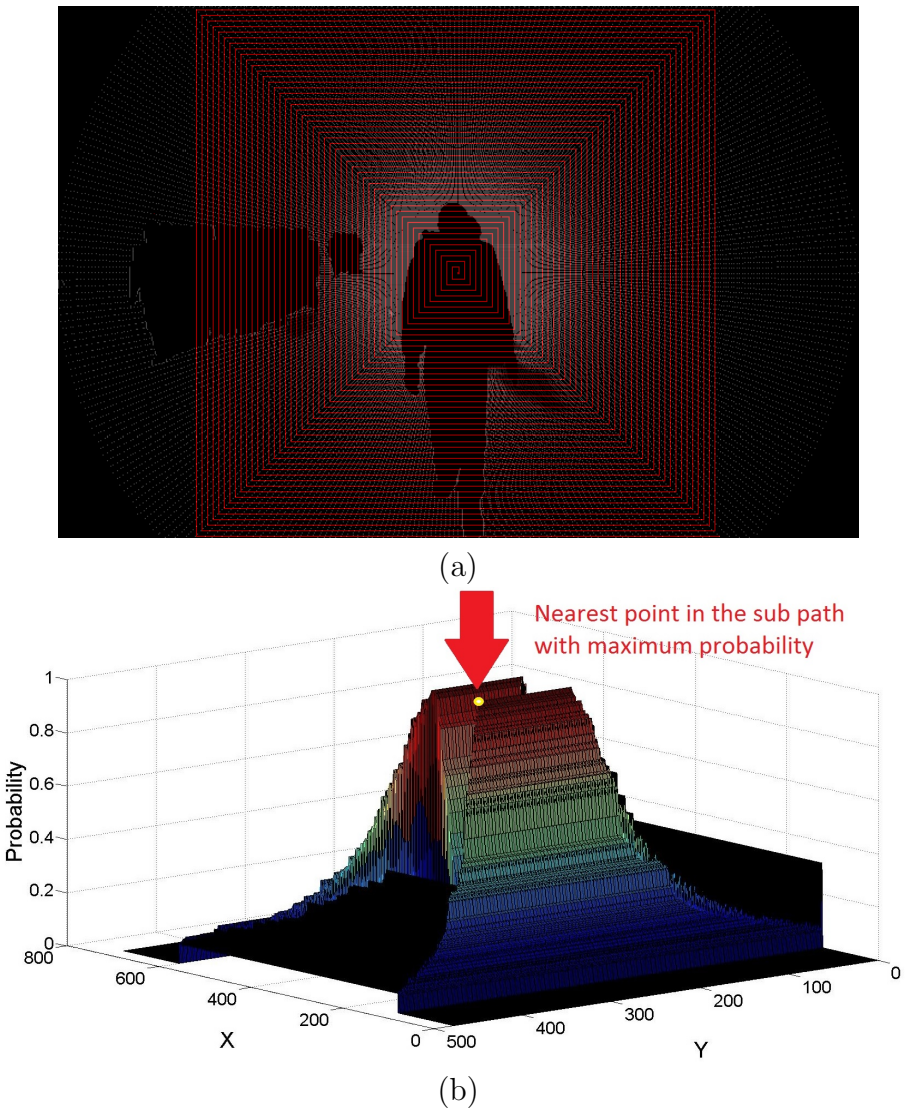
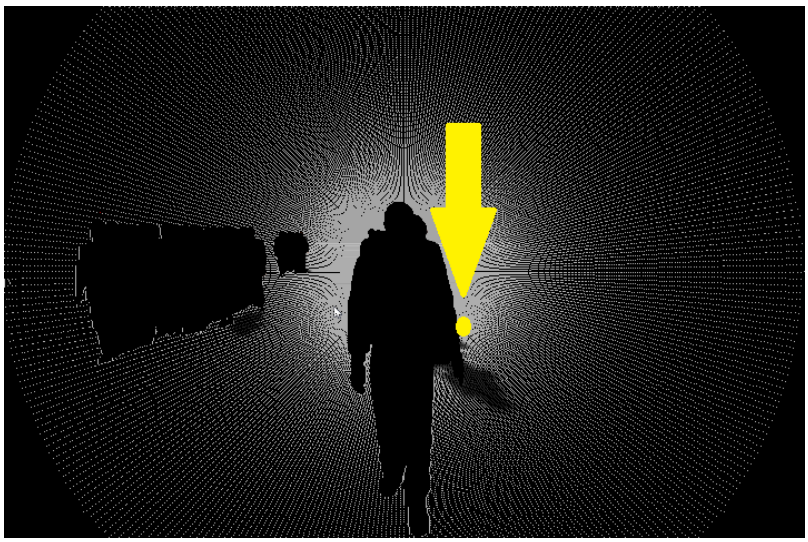
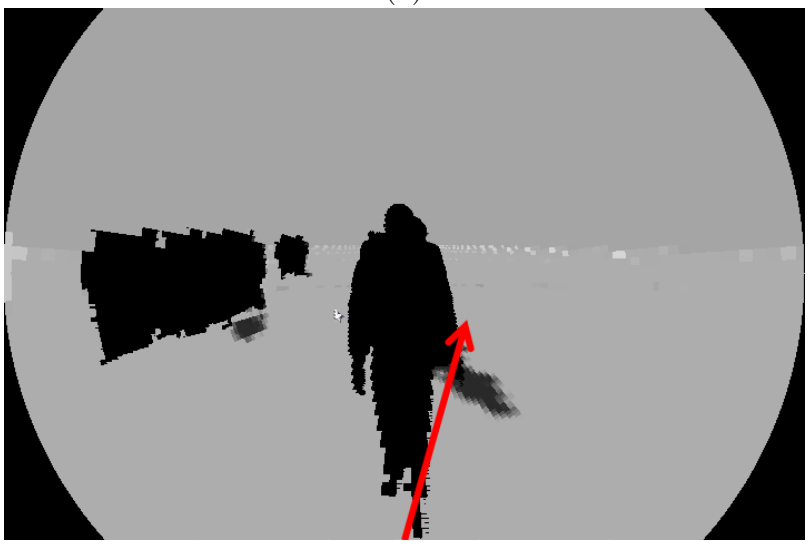


Figure 4.6: (a) Center-surround search path. Red line is a sub-path, which is sparse here for visualization purpose. (b) The generated 3d probability map. The yellow point represents the nearest point in the sub path with maximum probability which is the destination point. X and Y axes are width and length in image coordinate and the unit is pixels.

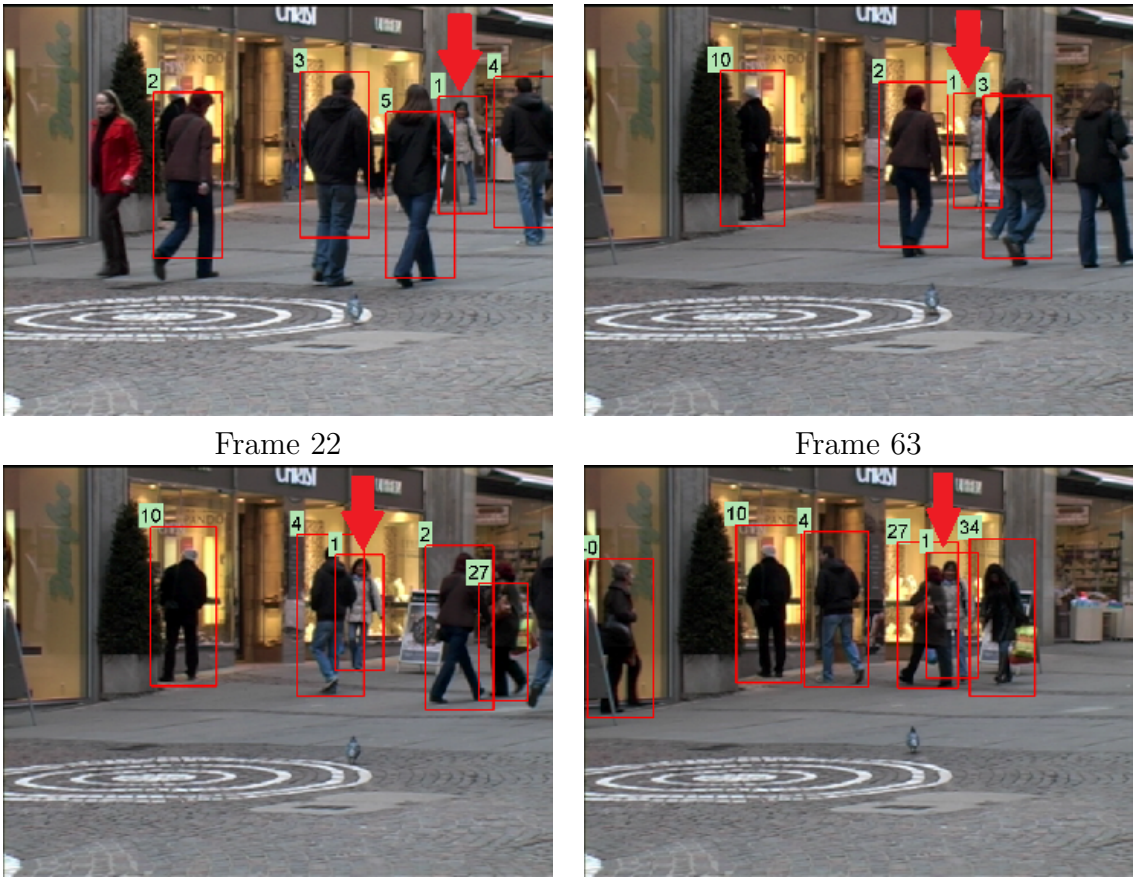


(a)



(b)

Figure 4.7: (a) Potential destination point in first-perspective view image. (b) The calculated moving angle based on attentive vision.



Frame 22 Frame 63

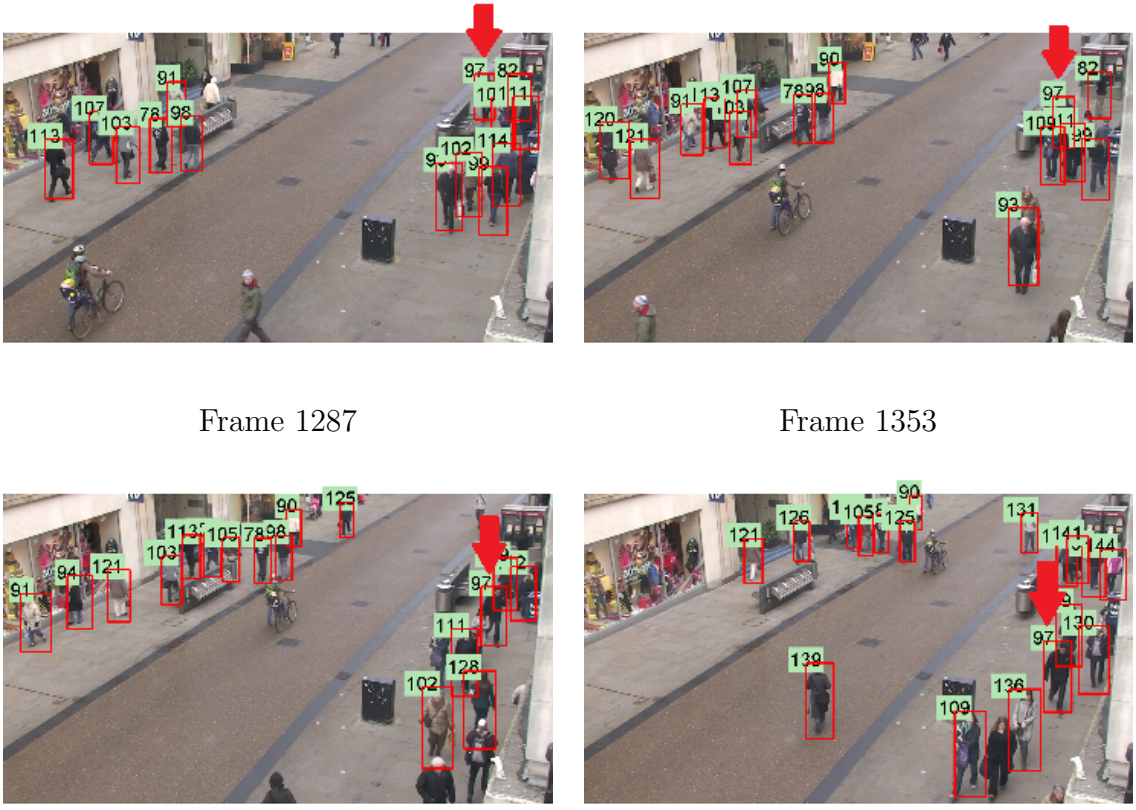
Frame 107 Frame 165

Figure 4.8: Tracking results of our approach on TUD statmitte dataset. Tracker under heavy occlusion and interaction: Object 1 is tracked correctly.

Frame 252 Frame 418

Frame 539 Frame 635

Figure 4.9: Tracking results of our approach on scenario one of TownCentre dataset. Long-term tracking under full occlusion, abrupt motion change and miss detection: Object 26 is tracked correctly in spite of significant change of motion direction.



Frame 1287 Frame 1353

Frame 1418 Frame 1509

Figure 4.10: Tracking results of our approach on scenario two of TownCentre dataset. Robust tracking in densely populated regions: Object 97 change the motion paths frequently due to the oncoming crowd.

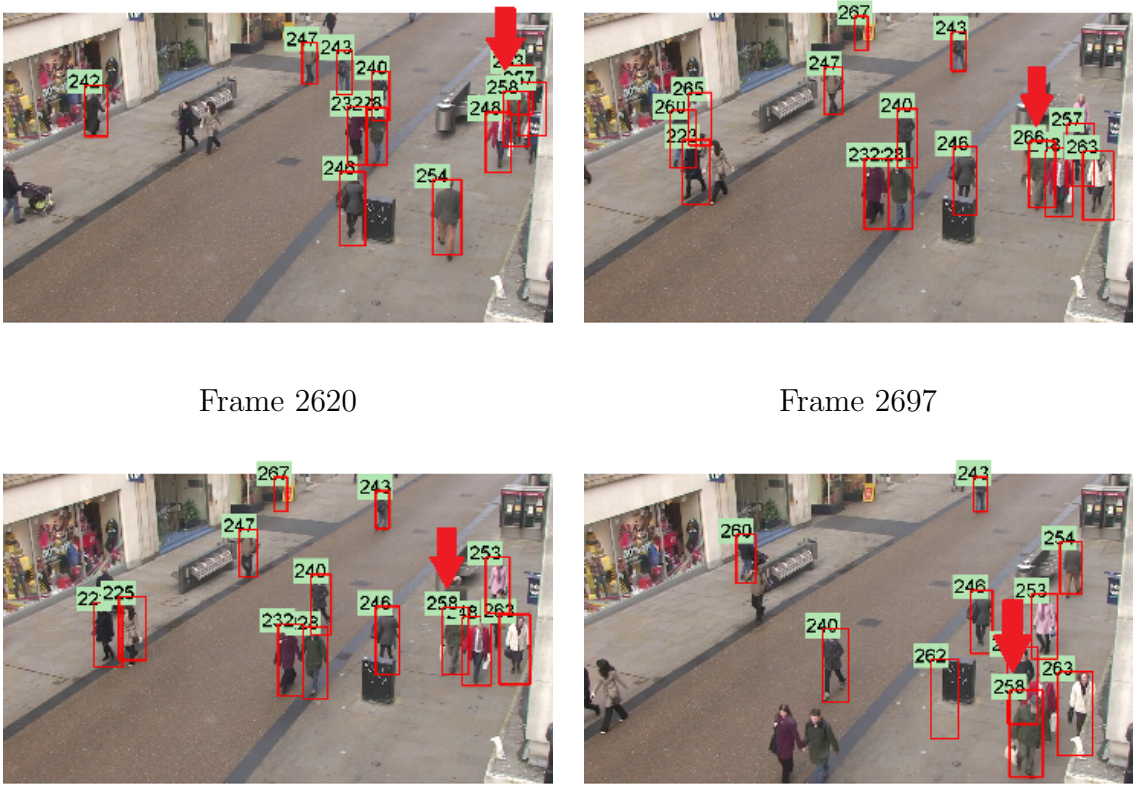


Figure 4.11: Tracking results of our approach on scenario three of TownCentre dataset. ID fragment correction: Object 258 suffers from ID fragment (but not ID switch) which is corrected in Frame 2703.

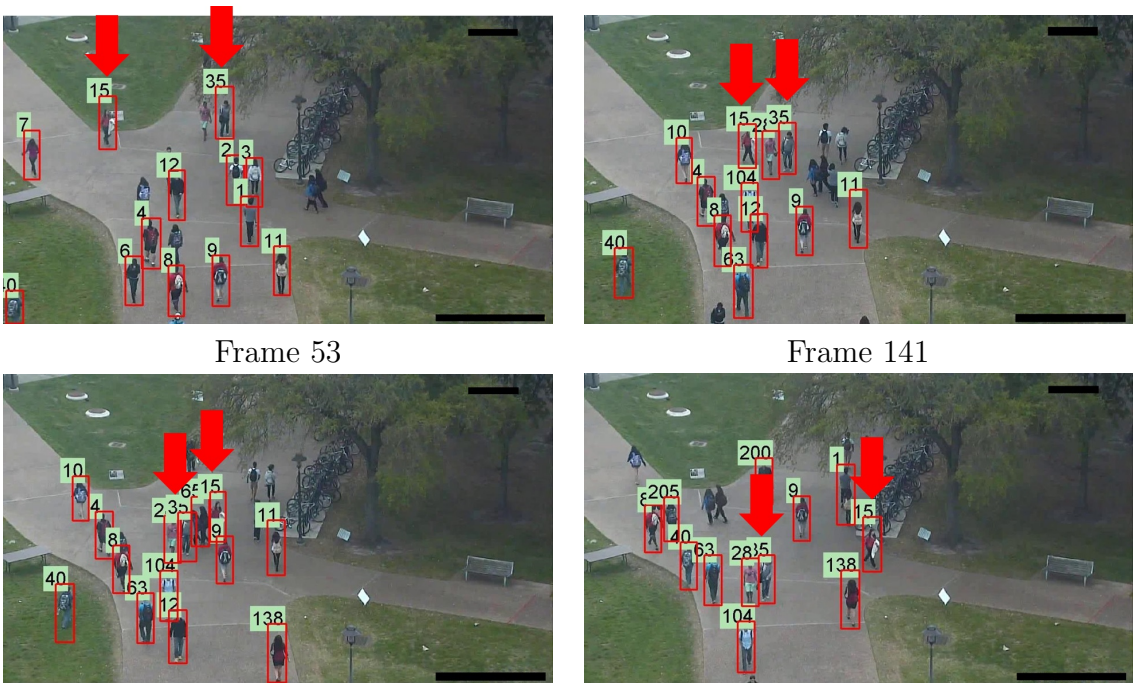


Figure 4.12: Tracking results of our approach on scenario two of our dataset. Attention vision prediction: Object 15 distracted from large amount of moving subjects which is corrected predicted by attentive vision modeling and recovered in Frame 201 and Frame 295.

Chapter 5

Hierarchical Group Structures in Multi-Person Tracking

5.1 Structure Preserving Object Tracking (SPOT)

Before introducing the group structure preserving object tracking, we present the structure preserving object tracking briefly [77]. Given a starting frame, each tracking target i ($i = 1, \ldots, N$) is represented by a bounding box denoted by $B_i = \{l_i, w_i, h_i\}$ with center location l_i and fixed width w_i and fixed height h_i. The tracking targets in the scene are denoted as $B = \{B_1, \ldots, B_N\}$. $\phi^b(I; B_i)$ denotes the feature vector for target i extracted from image $\mathbf{I}$.

SPOT defines a graph $G = (V, E)$ over all the targets with $V = \{B_1, \ldots, B_N\}$ and E represents the set of edges among all the targets. The multiple object tracking

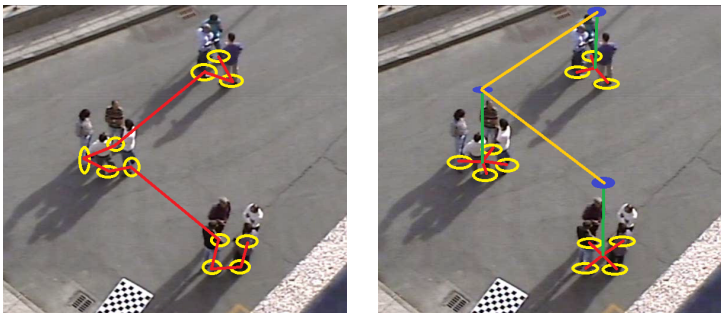
problem is defined as finding the best *configuration* $C = \{B_1^*, \ldots, B_N^*\}$ with highest score over the graph G as shown in Fig. 5.1(c). The problem is translated into an optimization problem: $\arg\max_{\{C\}} S(C; I, \theta)$ with

$$\begin{aligned} S(C; I, \theta) &= \sum_{i \in V} w_i^T \phi^b(I; B_i) \\ &\quad - \sum_{(i,j) \in E} \lambda_{ij} |(l_i - l_j) - e_{ij}|^2 \end{aligned} \tag{5.1}$$

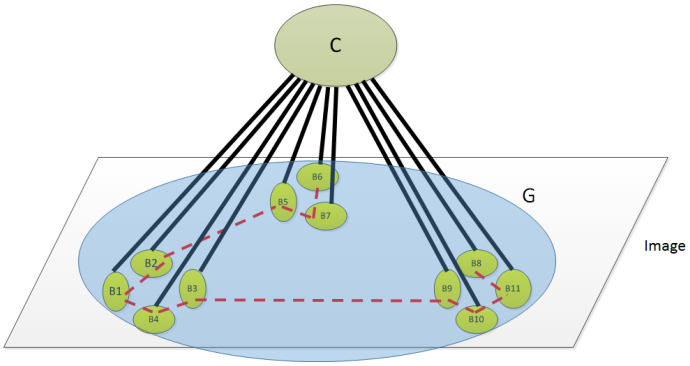
Herein, the parameters w_i represent linear weights on object features, e_{ij} represent the length and direction of the springs between object i and j, and θ denotes the set of all parameters including w and e. The parameter λ_{ij} is treated as a hyper-parameter.

5.2 Group Structure Preserving Object Tracking (GSPOT)

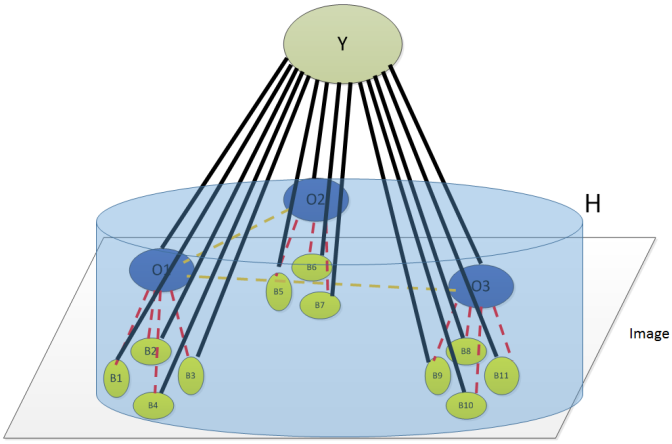
As shown in Fig. 5.1(a), SPOT links all the objects in the scene in a flat graph in the form of a minimum spanning tree. This inherently limits the ability to treat inter-group and intra-group dynamics uniquely. To explicitly incorporate the different levels of motion dynamics, the group structure preserving object tracking is built beyond structure-preserving object tracker (SPOT) by extending the single layer graph structure of SPOT to a two-layer graph structure. The structure of GSPOT is shown in Fig. 5.1(b), which unifies the inter-group and intra-group structure in a more natural way.



(a) SPOT Structure (b) GSPOT Structure



(c) SPOT Configuration



(d) GSPOT Configuration

Figure 5.1: Examples of graph structures and configurations created by SPOT and GSPOT. The SPOT configuration models both inter- and intra-group relationships using a single layer graph while the proposed GSPOT configuration treats inter- and intra-group relationships through separate hierarchies of the constructed graph.

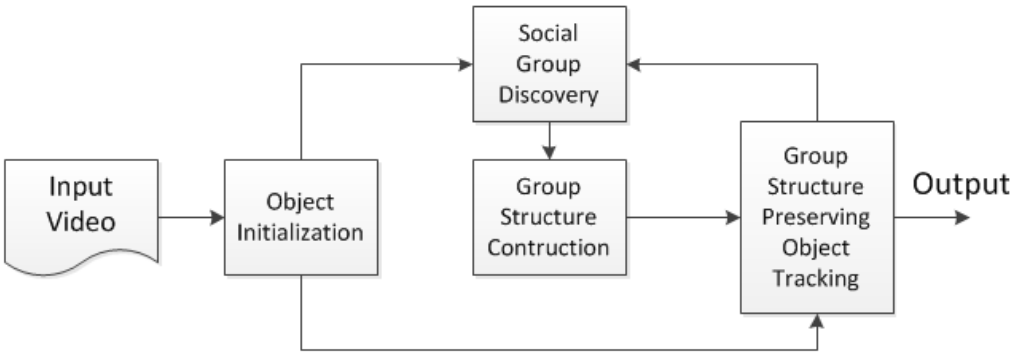


Figure 5.2: Depiction of the system framework.

5.2.1 System Framework

The system framework is illustrated in Fig. 5.2. The proposed method starts by getting the objects' initial positions, which are also used by the social group discovery component to obtain the spatial structures. With the obtained social groups, we construct the corresponding group structure and GSPOT uses the structure to track the objects robustly. Tracking evidences are collected and the social groups are updated intermittently so that the overall group structure can be updated accordingly.

5.2.2 Problem Formulation

Given the subject bounding boxes B_i and the respective feature vectors $\phi^b(I; B_i)$ as defined in section 5.1, we further define each social group k ($k = 1, \ldots, M$) in the scene as g_k. The group g_k includes one group center o_k and a target-group mapping

set $\Psi_k = \{\psi_{ki}\}$ $(i = 1, \ldots, N)$, where ψ is a target-group mapping given as:

$$\psi_{ki} = \begin{cases} 1 & \text{if target } i \text{ is mapped to group } k, \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

We include an added constraint $\sum_{k=1}^{M} \psi_{ki} = 1$ on the target-group mapping set and $\phi^g(I; g_k)$ denotes the feature vector for a group.

A graph $H = (G, O, \mathrm{T})$ is defined over all the targets and social groups. Note here, G is abused to denote all the inter-group graph structure $G = \{G_1, \ldots, G_k, \ldots, G_M\}$. T denotes the set of edges $\{\tau\}$ between group centers $O = \{o_1, \ldots, o_k, \ldots, o_M\}$. For each social group g_k, a sub-graph is defined as $G_k = (V_k, E_k)$ with $V_k = \{(B|\Psi_k = 1)\}$ and E_k represents the set of edges within a group. Within the defined structured, the multiple object tracking problem is defined as finding the best *configuration* over the whole graph $Y = \{B_1^*, \ldots, B_N^*, o_1^*, \ldots, o_M^*\}$ as shown in Fig. 5.1(d). Subsequently, we define the score of a configuration as:

$$\begin{aligned} S(Y; \mathbf{I}, \Theta) = & \sum_{k=1}^{M} w_k^T \phi^g(I; g_k) - \\ & \sum_{(p,q) \in L} \beta_{pq} |(c_p - c_q) - \tau_{pq}|^2 + \\ & \sum_{k=1}^{M} \left(\sum_{i \in V_k} w_i^T \phi^b(I; B_i) - \right. \\ & \left. \sum_{(i,j) \in E_k} \alpha_{ij} |(l_i - l_j) - e_{ij}|^2 \right) \end{aligned} \tag{5.3}$$

Herein, the parameters w_k represent linear weights on the group features, and e_{ij} and τ_{pq} represent the length and direction of the springs between inter-groups and intra-groups, respectively. The parameters α_{ij} and β_{pq} are treated as hyper-parameter

and are denoted as $\forall i, j : \alpha_{ij} = \alpha$ and $\forall p, q : \beta_{pq} = \beta$. The set of all parameters including w_k, w_i, τ_{pq}, and e_{ij} is denoted as Θ. In this dissertation, we define the group feature as the concatenation of target features in one group, which is denoted as $\phi^g(I; g_k) = \sum_{i \in V_k} \phi^b(I; B_i)$.

Group Structure and Inference. The inference of the model amounts to maximizing Eq. 5.3 over Y. Solving a complete connected graph is intractable. With the tree structured graph, we can solve the optimization by dynamic programming [23]. To make the inference tractable, we solve the optimization over the two-layer tree-structured graph H and $\{G_1, \ldots, G_k, \ldots, G_M\}$. We use two variants of the tree structure graph: (1) the star model based tree structure for G_k; and (2) a minimum spanning tree structure for H.

5.2.3 Parameter Learning and Iterative Parameter Learning.

After solving an optimal object configuration Y, we update the parameters by minimizing the structured SVM loss [66]:

$$\ell(\Theta; I, Y) = \max_{\hat{Y}} \left(s(\hat{Y}; I, \Theta) - s(Y; I, \Theta) + \Delta(Y, \hat{Y}) \right), \tag{5.4}$$

where $\Delta(Y, \hat{Y})$ is defined as:

$$\Delta(Y, \hat{Y}) = \sum_{k=1}^{M} \left(1 - \frac{V_k \cap \hat{V}_K}{V_k \cup \hat{V}_K} \right). \tag{5.5}$$

The loss function can be reshaped as:

$$\begin{aligned}\ell(\Theta; I, Y) \;=\; & \max_{\hat{Y}} \Big(vec(\Theta)^T \left(\hat{\Upsilon} - \Upsilon \right) - \\ & \sum_{k=1}^{M} \sum_{(i,j)\in E_k} \alpha \left(|\hat{m}_{ij}|^2 - |m_{ij}|^2 \right) - \\ & \sum_{(p,q)\in L} \beta \left(|\hat{n}_{pq}|^2 - |n_{pq}|^2 \right) + \Delta(Y, \hat{Y}) \Big),\end{aligned} \tag{5.6}$$

where $m_{ij} = l_i - l_j$, $n_{pq} = c_p - c_q$, and $\Upsilon = [\phi_1^g, \ldots, \phi_M^g, 2\alpha m_{i_1 j_1}, \ldots, 2\alpha m_{i_{|E_1|} j_{|E_1|}}, \ldots, 2\alpha m_{i_1 j_1}, \ldots, 2\alpha m_{i_{|E_M|} j_{|E_M|}}, 2\beta n_{p_1 q_1}, \ldots, 2\beta n_{p_{|L|} q_{|L|}}]^T$, $vec(\cdot)$ concatenates all parameters in a column vector.

We present two different methods to update the parameters. The first one is similar to the one used by Zhang *at al.* [77] and is given as:

$$\Theta \leftarrow \Theta - \frac{\ell(\Theta; I, Y)}{|\nabla_\Theta \ell(\Theta; I, Y)|^2 + \frac{1}{2K}} \nabla_\Theta \ell(\Theta; I, Y), \tag{5.7}$$

where we update all the parameters at the same time.

The second method updates the parameters in an iterative manner. We split the parameter set Θ as the union of Θ_1 and Θ_2. Θ_1 represents the intra-group feature parameter set including w_k and τ_{pq} and Θ_2 represents the inter-group feature parameter set including w_i and e_{ij}. The updates are also split into two stages. In the first stage only Θ_1 is updated as:

$$\Theta_1 \leftarrow \Theta_1 - \frac{\ell(\Theta; I, Y)}{|\nabla_\Theta \ell(\Theta; I, Y)|^2 + \frac{1}{2K}} \nabla_\Theta \ell(\Theta; I, Y). \tag{5.8}$$

After Θ_1 is updated, one more optimization (solve the Eq. 5.3) is performed prior

to updating Θ_2 as the second stage. Θ_1 is kept fixed and Θ_2 is updated as:

$$\Theta_2 \leftarrow \Theta_2 - \frac{\ell(\Theta; I, Y)}{|\nabla_\Theta \ell(\Theta; I, Y)|^2 + \frac{1}{2K}} \nabla_\Theta \ell(\Theta; I, Y). \tag{5.9}$$

To refer to the joint parameter optimization approach as "GSPOT" and the iterative approach as "iGSPOT".

5.3 Experiments

To evaluate the merit of our proposed hierarchical-structure-based model, we conducted experiments on several public datasets to compare the performance of the proposed methods against existing methods as well as to evaluate the merits of different components of our approach. The "S15-FM" sequence was included from the "Friends Meet" dataset [10]. It shows multiple occurrences of persons merging and splitting from a group and is the most challenging sequence in the whole dataset. One video sequence was included from the "BEHAVE" Interactions Test Case Scenarios [46], which includes the scenario showing two groups merging into a larger group. One video sequence from the "QIL" dataset was also included [70] in which several individual persons merge into a group and move together. Finally, one video sequence was included from the "Crowd by Example" dataset that has poor image quality and shows several groups moving around in a natural manner. The tracking performance is measured based on CLEAR MOT metrics [13]. We report multiple object tracking precision (MOTP), miss rate (MISS), false positive rate (FP), number of ID switches (IDS) and multiple object tracking accuracy (MOTA). The threshold for building a matched pair between a tracking result and the ground truth is selected as half of the

bounding rectangles' diagonal in the ground truth. It should be noted that MOTP measures the ability of a tracker to estimate precise pedestrian positions, which is independent of an algorithm's tracking accuracy. MOTP is computed as the average error of center position of matched pairs' over all frames, measured in pixels. Note here, the MOTA result could be a negative number since it is computed as:

$$MOTA = 1 - \frac{\#\text{ of } miss + \#\text{ of } fp + IDswitches}{\#\text{ of } groundtruth}. \tag{5.10}$$

5.3.1 Evaluation: Social Group Discovery

We explore three methods to discover the social groups within a scene. The first two methods are detailed in [26] and [64], respectively. Method described by Ge *at al.* [26] uses both spatial and temporal cues to group subjects based on their trajectories while the method by Khai *et al.* [64] leverages the social cues from a single frame based on subject position and pose to find dominant groups. The third method we use is based on k-means clustering [59]. The number of clusters or groups, k, is set to be the same as the number of groups identified by the first method [26]. We integrated the three methods into "GSPOT" with all the other component being identical and evaluated the tracking performance on the four video sequences. The average MOTA performance in Table 5.1 shows that the social grouping defined in [26] gives best overall performance, which could be explained by the advantage of cumulative spatio-temporal evidences. This metric is based on an agglomerative clustering where the group membership ρ_{ij} denotes the number of frames in which pedestrian i and j are in same group. There is a link between two pedestrians i and

GSPOT	MOTP	MOTA	FP	FN	IDS
K-means [59]	5.58	70.57%	13.56%	13.56%	29
Khai *et al.* [64]	4.96	75.98%	11.94%	11.94%	7
Ge *at al.* [26]	**4.57**	**99.72**%	**0.14**%	**0.14**%	**0**

Table 5.1: Average CLEAR MOT evaluation results on three grouping methods. The best results are in bold.

j if $p_{ij} > \tau$. We set $\tau = 10$ [26]. ω_{ij} represents the average pairwise distance over all the frames when pedestrians i and j are in same group. A modified Hausdorff distance $H(A, B)$ is derived from pairwise distance matrix and is used to measure inter-group closeness between groups A and B as $H(A, B) = (h(A, B) + h(B, A))/2$, where

$$h(A, B) = \frac{\sum_{i=1}^{|A|} \sum_{l=1}^{\lceil |B|/2 \rceil} w_{ij}^{l}}{|A| \times \lceil |B|/2 \rceil} \tag{5.11}$$

and w_{ij}^{l} is the l^{th} smallest distance among all the distances, which are derived between pedestrian i in group A and all the pedestrians $j \in B$. The social group is then obtained by agglomerative clustering. The merge step is governed by pairwise group Hausdorff distance and the merging is stopped by intra-group tightness criterion shown in Eq. 5.12.

$$e_{A+B} < \hat{e}_{|A+B|} + (e_A - \hat{e}_{|A|} + e_B - \hat{e}_{|B|}) \tag{5.12}$$

where e_A, e_B, e_{A+B} are the total number of links in group A, group B, and the merged group $A + B$, respectively. $\hat{e}_A$, $\hat{e}_B$, $\hat{e}_{A+B}$ denote the minimal expected number of links in group A, group B, and merged group $A + B$, respectively. For the remaining results shown in the dissertation, method by Ge *at al.* is set as the default social group discovery method.

5.3.2 Evaluation: Group Structure

We present variants on updating the identified group structures based on the discovered social groups at initialization to evaluate the merit of the hierarchical graph in GSPOT. The inter-group structure is initialized but not updated in "GSPOT.v1" while only the intra-group structure is updated during tracking. The intra-group structure is initialized in "GSPOT.v2" but not updated while only the inter-group structure is updated during tracking. Note here that the parameters of the structures are reinitialized after new group discovery for "GSPOT.v1" and "GSPOT.v2" although there is no update in each grouping interval for the component held constant. From the tracking performance presented in Table 5.2, it shows that (1) both precise inter-group and intra-group structures are critical without which the tracking performance degraded in all of scenarios; (2) intra-group structure update has a significant impact over inter-group structure update in tracking performance; and (3) inter-group structure update contributes to the tracking performance when the structure in individual groups changes significantly.

5.3.3 Evaluation: Parameter Learning

To fully evaluate the benefit of iterative parameter learning, we compared the performance of iGSPOT against GSPOT. Besides the original trackers, two variant are evaluated as well. iGSPOT.g1 and GSPOT.g1 denote no group structure updates after initialization in the first frame with the rest of components being the same. In the second variant, the default social grouping method in both iGSPOT and GSPOT

BEHAVE	MOTP	MOTA	FP	FN	IDS
GSPOT.v1	10.86	99.94%	0.03%	0.03%	**0**
GSPOT.v2	9.64	21.41%	39.28%	39.28%	1
GSPOT	**5.85**	**100.00**%	**0.00**%	**0.00**%	**0**
QIL	MOTP	MOTA	FP	FN	IDS
GSPOT.v1	4.53	71.84%	14.08%	14.08%	**0**
GSPOT.v2	**4.31**	44.82%	27.59%	27.59%	**0**
GSPOT	5.03	**100.00**%	**0.00**%	**0.00**%	**0**
FM	MOTP	MOTA	FP	FN	IDS
GSPOT.v1	5.88	99.50%	0.25%	0.25%	**0**
GSPOT.v2	5.42	58.14%	20.93%	20.93%	**0**
GSPOT	**4.03**	**99.80**%	**0.10**%	**0.10**%	**0**
CROWD	MOTP	MOTA	FP	FN	IDS
GSPOT.v1	4.43	74.53%	8.83%	8.83%	100
GSPOT.v2	4.27	22.66%	38.52%	38.52%	4
GSPOT	**3.39**	**99.06**%	**0.47**%	**0.47**%	**0**

Table 5.2: CLEAR MOT evaluation results on four datasets. The best results are in bold.

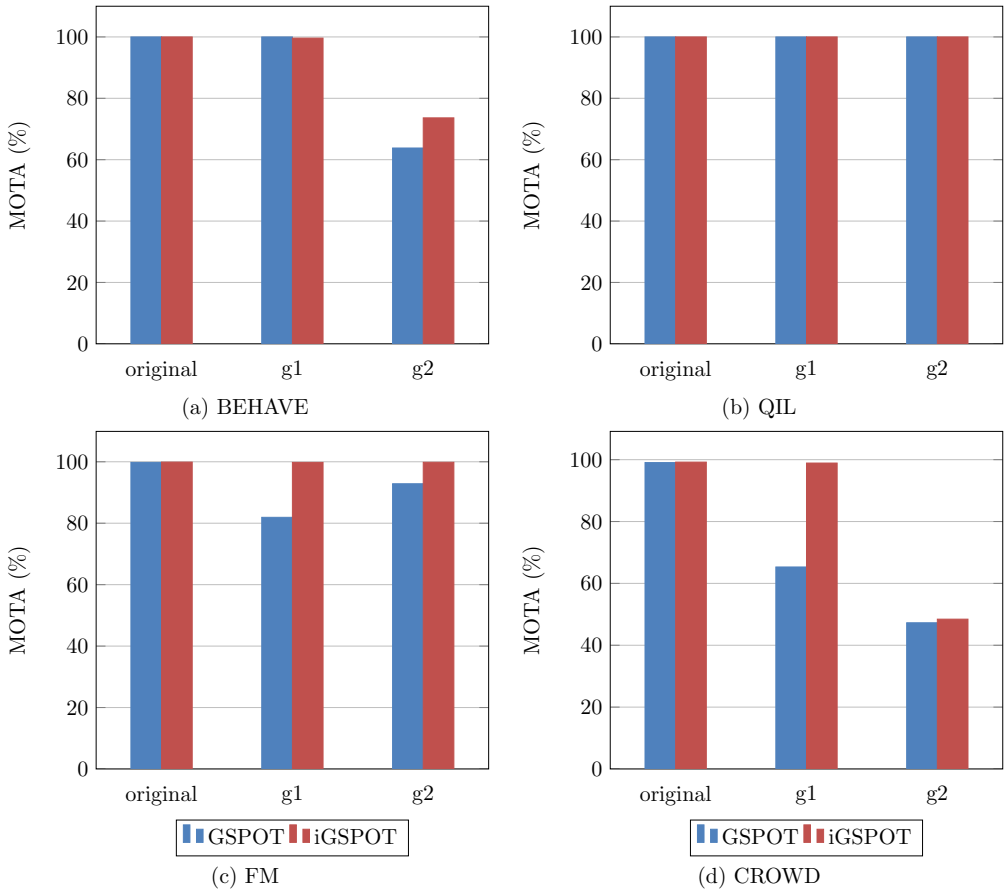


Figure 5.3: Evaluation on the effects of GSPOT, iGSPOT and their variants.

are replaced by the method of Khai *et al.* [64]. We denote them as iGSPOT.g2 and GSPOT.g2. The tracking accuracy is assessed to measure performance improvement as shown in Fig. 5.3. As seen, the iterative parameter update gives more robust performance under various settings.

5.3.4 Evaluation: Overall performance

To evaluate the overall tracking performance, we compared performance results of the proposed tracker against four competing approaches for all the video sequences. The trackers compared included: (1) Multiple Instance Learning Tracker [8] (MIL), (2) Structure preserving object tracker [77] (SPOT), (3) Multiple structure preserving object tracker with social group discovery (MSPOT.v1), and (4) Multiple structure preserving object tracker with fixed grouping (MSPOT.v2). To compare fairly, we implemented MIL tracker with HOG feature and the parameters of MIL are tuned to get the best results. SPOT was run using the implementations provided by [77]. MSPOT by name is an extension of SPOT that runs multiple SPOT trackers on different groups. MSPOT.v1 leverages the same social group discovery method as used in our model and is updated in the tracking process. MSPOT.v2 uses the fixed group setting that is set manually. The performance of the five trackers on all four datasets is presented in Table 5.3. The results in the table show that (1) our iGSPOT and GSPOT tracker outperform the state-of-the-art tracker SPOT and MIL indicating that hierarchical structure contributes towards improved tracking; (2) MSPOT treats objects in the scene and builds multiple independent structure for each one resulting in improved performance compared to SPOT in BEHAVE and FM datasets but contributes to worse performance in the other two datasets; (3) iGSPOT and GSPOT outperform the two MSPOT trackers, which suggests that the structure between each group pairs should not be overlooked; (4) MIL results in worse performance compared with SPOT in three of the sequences demonstrating that incorporation of spatial structures aids tracking.

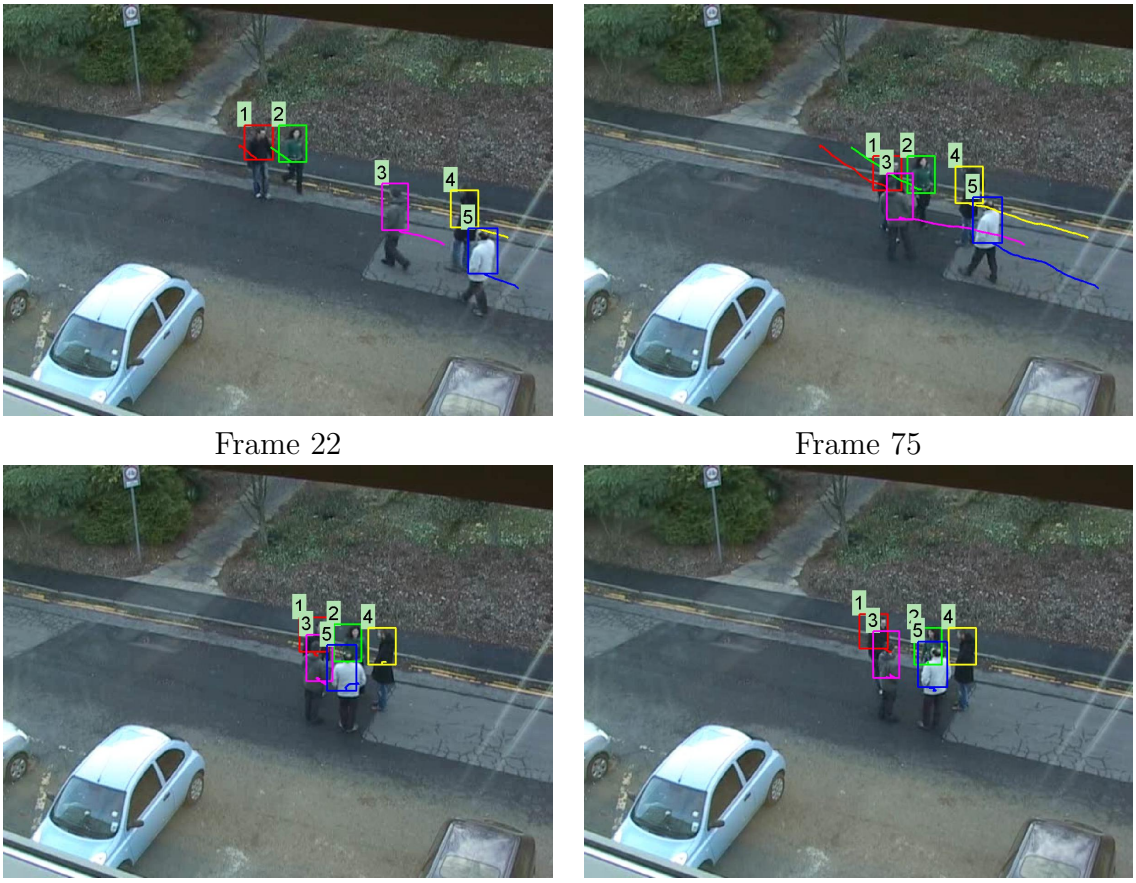
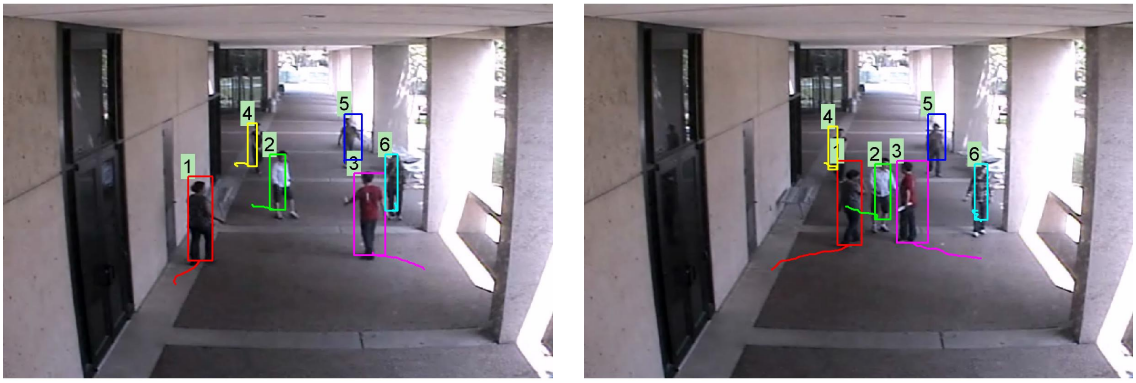


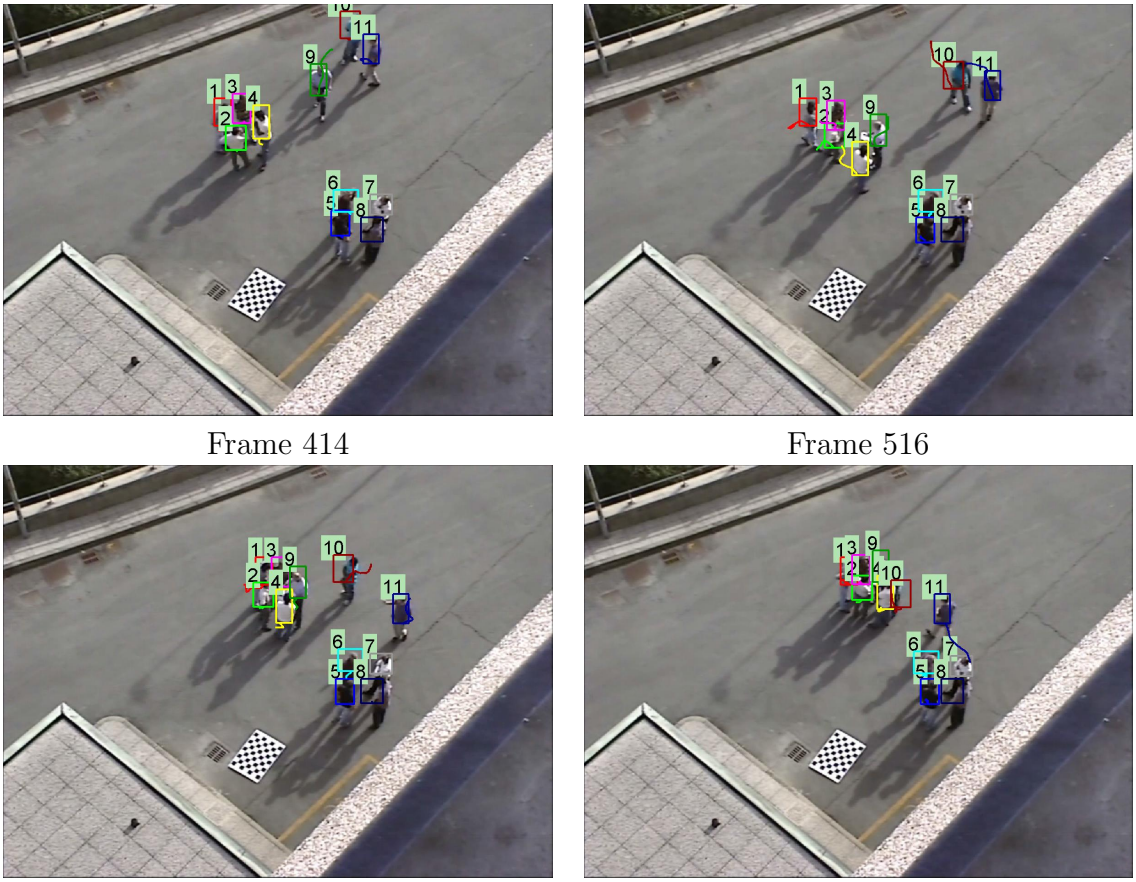
Figure 5.4: Tracking results of our approach on BEHAVE dataset.



Frame 67 Frame 126

Frame 263 Frame 356

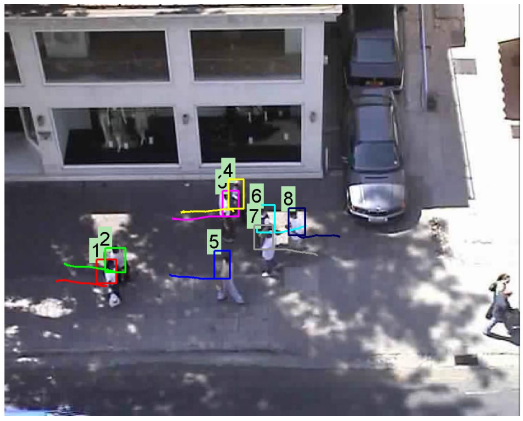
Figure 5.5: Tracking results of our approach on QIL dataset.



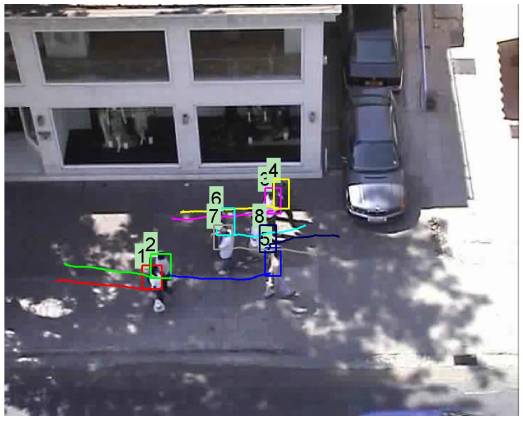
Frame 414 Frame 516

Frame 635 Frame 830

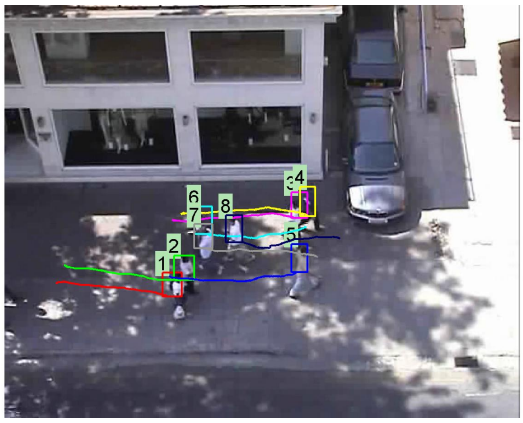
Figure 5.6: Tracking results of our approach on FM dataset.



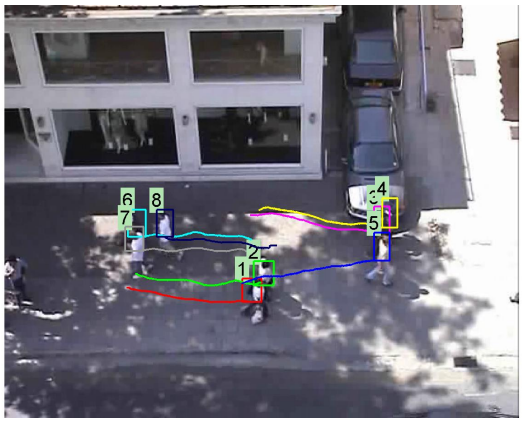
Frame 55



Frame 84



Frame 99



Frame 147

Figure 5.7: Tracking results of our approach on CROWD dataset.

BEHAVE	MOTP	MOTA	FP	FN	IDS
MIL	12.18	-27.41%	63.03%	63.03%	43
SPOT	16.62	16.81%	41.59%	41.59%	**0**
MSPOT.v1	7.69	22.31%	37.84%	37.84%	64
MSPOT.v2	7.97	57.87%	21.06%	21.06%	**0**
GSPOT	**5.85**	**100.00**%	**0.00**%	**0.00**%	**0**
iGSPOT	**5.85**	**100.00**%	**0.00**%	**0.00**%	**0**
QIL	MOTP	MOTA	FP	FN	IDS
MIL	7.76	26.49%	34.17%	34.17%	118
SPOT	**4.53**	71.84%	14.08%	14.08%	**0**
MSPOT.v1	14.12	-6.89%	51.75%	51.75%	77
MSPOT.v2	7.51	28.46%	31.14%	31.14%	211
GSPOT	5.03	**100.00**%	**0.00**%	**0.00**%	**0**
iGSPOT	5.03	**100.00**%	**0.00**%	**0.00**%	**0**
FM	MOTP	MOTA	FP	FN	IDS
MIL	7.05	26.77 %	31.25%	31.25%	636
SPOT	6.36	15.01 %	36.90%	36.90%	663
MSPOT.v1	4.21	60.82%	17.83%	17.83%	209
MSPOT.v2	5.85	59.15%	18.57%	18.57%	220
GSPOT	**4.03**	99.80%	0.10%	0.10%	**0**
iGSPOT	4.05	**99.90**%	**0.05**%	**0.05**%	**0**
CROWD	MOTP	MOTA	FP	FN	IDS
MIL	6.58	37.89%	28.52%	28.52%	65
SPOT	3.89	56.72%	21.56%	21.56%	2
MSPOT.v1	5.63	2.66%	47.27%	47.27%	36
MSPOT.v2	3.60	30.16%	34.38%	34.38%	14
GSPOT	**3.39**	99.06%	0.47%	0.47%	**0**
iGSPOT	3.40	**99.22**%	**039**%	**0.39**%	**0**

Table 5.3: CLEAR MOT evaluation results on four datasets. The best results are in bold.

Chapter 6

Social Interaction based Tracking

The social force model by Helbing [33] is a computational model in which the interactions among pedestrians are described by using the concept of forces between physical entities. Each pedestrian feels a social force from other pedestrians that is proportional to the distance between them. In this model, a pedestrian $i = 1, \ldots, H$ makes motion decisions based on the sum of forces $\mathbf{F}_i$ exerted. Under the modeled social force, the motion model that predicts the positional information for a tracked pedestrian i is given by:

$$\frac{\mathbf{F}_i}{m_i} = \frac{\partial \mathbf{v}_i}{\partial t}, \tag{6.1}$$

where $\mathbf{v}_i$ is the instantaneous velocity and m_i is the mass.

With unknown and complicated social interaction in the scene, we aim to break social interaction into combination of atomic social effects and quantify them with social force model similar as [33] and give informative prediction about human motion.

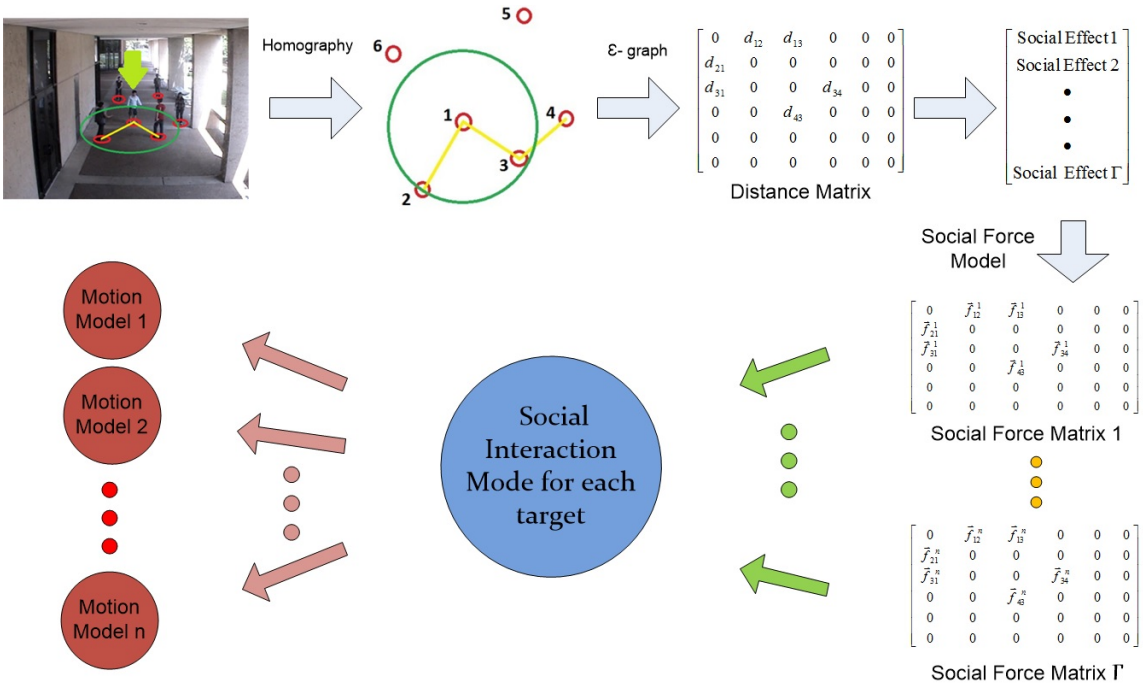


Figure 6.1: Depiction of the social interaction decomposition framework

6.1 Social Interaction Decomposition

Before decomposing the social interaction, we first define what social effect, social link, and social interaction mode are. Social effect by name is certain social interaction context that drives human motion behavior and can be defined arbitrarily (e.g. follow, spread, pass, repulsion, attraction). Social link between two persons represents one of the predefined social effects at the current time instance. Note here that the social link is directed. Social interaction mode for one pedestrian represents the set of social effects that happen over all social links. It is hard to know the exact social effect for every social link without an explicit knowledge of pedestrians' intent. The same difficulty holds for knowing the exact social interaction mode.

With the definition of social interaction mode, we treat pedestrian motion as the outcome of linear combination of potential social interaction modes with dynamically adjusted weights at different moments in time. We decompose unknown social interaction into multiple social interaction modes which are composed by predefined atomic social effects. The social effects is quantified using a social force model as shown in Fig. 6.1. Give the social links of one pedestrian $l = 1, \ldots, L$ and atomic social effects $\gamma = 1, \ldots, \Gamma$, the n^{th} social interaction mode is denoted as: $d_n = \{\gamma_1, \ldots, \gamma_L\}$. The total number of social interaction modes is $N = \Gamma^L$.

We drop the subscript of pedestrian i's social force with $\mathbf{F} = \mathbf{F}_i$ here for simplicity and it is given by:

$$\mathbf{F} = \sum_{n=1}^{N} \omega_n \Phi\left(d_n\right) = \sum_{n=1}^{N} \omega_n \cdot \left(\sum_{l=1}^{L} \phi(\gamma_l) \right) \tag{6.2}$$

where $\Phi(d_n)$ represents quantified social force under social interaction mode d_n, ω_n is a weighting coefficient, $\phi(\gamma_l)$ represents social force of social link l under social effect γ_l. $\mathbb{D}_i$ is the set of social interaction modes and $d_n \in \mathbb{D}_i$. We set $\phi(\gamma_l) = \phi_\gamma(l)$ for later explanation.

6.1.1 Atomic social effects and force model

In this dissertation, we mainly explore three atomic social effects: repulsion, attraction, and non-interaction. The underlying assumption is that all the other social interaction can be represented as the combination of these three effects. The atomic social effects can be replaced to fit some other specific scene context. The repulsion effect captures the behavior where people try to avoid collisions with each other and

the attraction effect captures the behavior when a person approaches another person with an intent to meet. Non-interaction by name means pedestrians are independent from each other. With these three atomic units, the social effects pool is instantiated as $\gamma = \{+, -, 0\}$.

We quantify the atomic social effect with attraction and repulsion forces model, which share a similar model with different directions and monotonicity. Similar as Helbing's model [33], attraction, $\phi_+(\cdot)$, and repulsion, $\phi_-(\cdot)$, force models of pedestrian i with social link $i \to j$ are denoted as:

$$\phi_+(i \to j) = F^a * e^{(\frac{d_{ij} - r_{ij}}{b})}\mathbf{u}_{ij}, \tag{6.3}$$

and

$$\phi_-(i \to j) = F^r * e^{(\frac{r_{ij} - d_{ij}}{b})}\mathbf{u}_{ji}, \tag{6.4}$$

where F^r and F^a are the magnitudes of repulsion and attraction force, respectively, b is the boundary of the influence of the force, d_{ij} is the Euclidean distance between i and j, $\mathbf{u}_{ji}$ is the unit vector from j to i, and $\mathbf{u}_{ij}$ is the unit vector from i to j. The private sphere of a pedestrian is represented by a circle of radius r with r_i and r_j defining the private sphere of pedestrians i and j, respectively. Further, $r_{ij} = r_i + r_j$ defines the radius of influence for pedestrians i and j, respectively. $\phi_0(\cdot) = 0$ denotes non-interaction effect.

Case Analysis: In the example shown in Fig. 6.2, the decomposition is defined by first building social links with other pedestrians from single pedestrian. Each social link is hypothesized to exhibit repulsion, attraction, or non-interaction effects, which are denoted by $\{+\}$, $\{-\}$, and $\{0\}$ respectively. The repulsion and

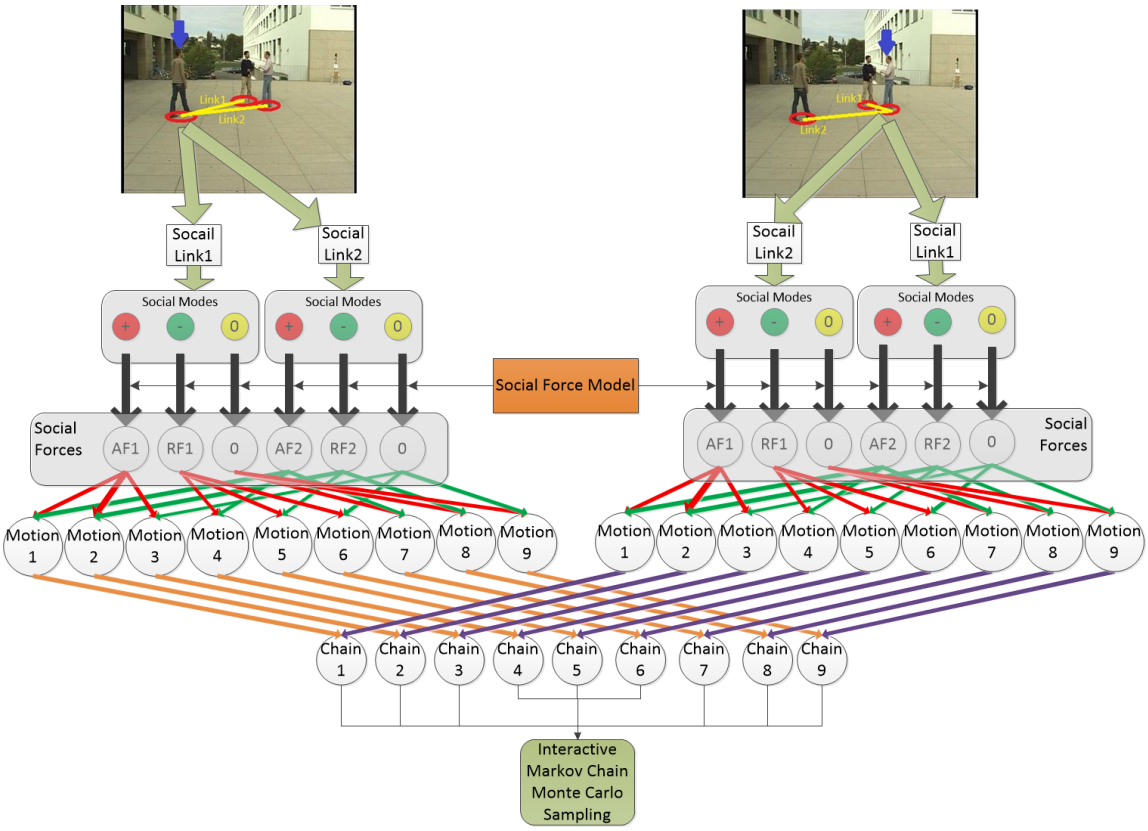


Figure 6.2: Example of the social interaction decomposition.

attraction effects are translated into social forces by Eq. 6.4 and Eq. 6.3 and $\{0\}$ is equal with no force. The set of potential social interaction modes $\mathbb{D}$ is composed by the social effects of each social link: $\{\{+,+\},\{+,-\},\{-,+\},\{-,-\},\{+,0\},\{0,-\},\{-,0\},\{0,-\},\{0,0\}\}$. A motion model is derived from the sum of social forces based on the social effects in one interaction mode. The maximum number of $|\mathbb{D}_i|$ is represented by N^{max}. $\{\mathbb{D}_i\}$ of each pedestrians are preprocessed to have the same size N^{max} by replicating the social interaction modes. This ensures that the same number of motion models are maintained for each pedestrian. The motion models are incorporated into Markov chains as shown in Fig. 6.2.

Within the context of a Bayesian tracking framework, the posterior distribution for each interaction mode d_n, given a motion prior, can be approximated by constructing a Markov chain for sampling. Hence, an increase in the number of social links leads to an increase of N^{max} and a corresponding increase in the number of Markov chains. Considering H pedestrians in a scene with three atomic social effects, if every pedestrian has a social interaction link with each other, then, the number of chains is given by $N^{max} = \Gamma^L = 3^{H-1}$. There is an exponential growth in the number of chains with respect to the number of pedestrians which is computationally unsustainable.

We address this issue by limiting the number of social links between pedestrians based on the distance between them. The construction of social links is based on the ϵ-graph [19]. A link is established between pedestrian i and j if $E(i,j) < \epsilon$ where E is the Euclidean distance in real world coordinates. By adjusting the value of ϵ, we build sparse social links among pedestrians. In this dissertation, ϵ is set to be equal to the value of the forces' boundary b which adjusts the sparseness of social links and is set empirically.

6.1.2 Motion Model with Social Interaction Modes

In a Bayesian context, the tracking problem is to quantify the posterior probability $p(x_t|y_{1:t})$, where the observations are specified by $y_{1:t} = \{y_1, y_2, ..., y_t\}$. Given the new observation y_t at time t, the posterior probability is estimated by:

$$p(x_t|y_{1:t}) = cp(y_t|x_t) \int p(x_t|x_{t-1}) p(x_{t-1}|y_{1:t-1}) dx_{t-1}, \tag{6.5}$$

where c is a normalization constant, $p(y_t|x_t)$ specifies the likelihood function of the current observation given the current state, $p(x_t|x_{t-1})$ specifies the probability of the current state given the previous state, and $p(x_{t-1}|y_{1:t-1})$ specifies the previous posterior probability.

The goal of our proposed method is to find the best state $\hat{x}_t$ at time t given the observation $y_{1:t}$. It can be obtained by using the Maximum a Posteriori (MAP) estimate over the M samples at each time t, denoted by:

$$\hat{x}_t = \arg\max_{x_t^\ell} p(x_t^\ell|y_{1:t}) \text{ for } \ell = 1, \ldots, M \ , \tag{6.6}$$

where x_t^ℓ indicates the ℓ^{th} sample of the state x_t. Our method estimates an accurate value of posterior probability by designing a sophisticated motion model $p(x_t|x_{t-1})$. Following the decomposition idea by Kwon and Lee [39], the motion model is designed as the weighted linear combination of its basic components:

$$p(x_t|x_{t-1}) = \sum_{\lambda=1}^{\Lambda} w_t^\lambda p_\lambda(x_t|x_{t-1}), \quad \text{and} \quad \sum_{\lambda=1}^{\Lambda} w_t^\lambda = 1, \tag{6.7}$$

where $p_\lambda(x_t|x_{t-1})$ denotes the λ^{th} basic motion model, w_t^λ is the weighting variable at time t and will be estimated implicitly in the interaction process of IMCMC as shown in Eq. 6.12, and Λ is the number of decomposed motion models, which is equal to cardinality of the social interaction mode set. Hence, $\Lambda = N^{max}$.

Let $\mathbf{c}_t = [\mathbf{c}_{1,t}, \mathbf{c}_{2,t}, \ldots, \mathbf{c}_{H,t}]^T$ and $\mathbf{v}_t = [\mathbf{v}_{1,t}, \mathbf{v}_{2,t}, \ldots, \mathbf{v}_{H,t}]^T$ denote the positions and velocities of H pedestrians, respectively, at time t. The position information for pedestrians in the next frame is predicted using the computed force denoted by $\mathbf{F}^\lambda$=[$\Phi_1(d_\lambda)$,$\Phi_2(d_\lambda)$, ...,$\Phi_H(d_\lambda)$]^T. Using a constant velocity motion model, the motion prediction is defined as $\mathbf{c}_t = \mathbf{c}_{t-1} + \mathbf{v}_{t-1}\Delta t$ 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:

$$\begin{bmatrix} \mathbf{c}_t^\lambda \\ \mathbf{v}_t^\lambda \end{bmatrix} = \begin{bmatrix} \mathbf{c}_{t-1} + \mathbf{v}_{t-1}\Delta t + \frac{1}{2}\frac{\mathbf{F}^\lambda}{m}\Delta t^2 \\ \mathbf{v}_{t-1} + \frac{\mathbf{F}^\lambda}{m}\Delta t \end{bmatrix}, \tag{6.8}$$

where $\mathbf{c}_t^\lambda$ and $\mathbf{v}_t^\lambda$ denote the predicted positions and velocity in the motion model λ of each pedestrians, respectively. For a single pedestrian i, the motion prediction $\mathbf{c}_{i,t}$ is given by a set of locations predicted by each of its social interaction modes, $\mathbf{c}_{i,t} = \{\mathbf{c}_{i,t}^1, \mathbf{c}_{i,t}^2, \ldots, \mathbf{c}_{i,t}^\Lambda\}$. In this work, each basic motion model $p_\lambda(x_t|x_{t-1})$ is modeled as a Gaussian distribution and is given by:

$$p_\lambda(x_t|x_{t-1}) \propto \mathcal{N}(\mathbf{c}_t; \mathbf{c}_t^\lambda, \sigma^2). \tag{6.9}$$

6.1.3 Observation Model

Following Perez *et al.*. [50], we use a color observation model based on the Hue-Saturation-Value (HSV) space. Given the initialization for any pedestrian to be tracked (bounding box), we perform a kernel density estimate, $g^* = g(x_0)$, of the color distribution in frame 0. The data likelihood is derived based on the Bhattacharyya similarity coefficient and is defined as $S[g^*, g(x_t)]$. The likelihood of observation, based on Gibbs distribution, is given by:

$$p(y_t|x_t) \propto \exp(-\tau S^2[g^*, g(x_t)]), \tag{6.10}$$

where $\tau = 20$ as suggested in [50].

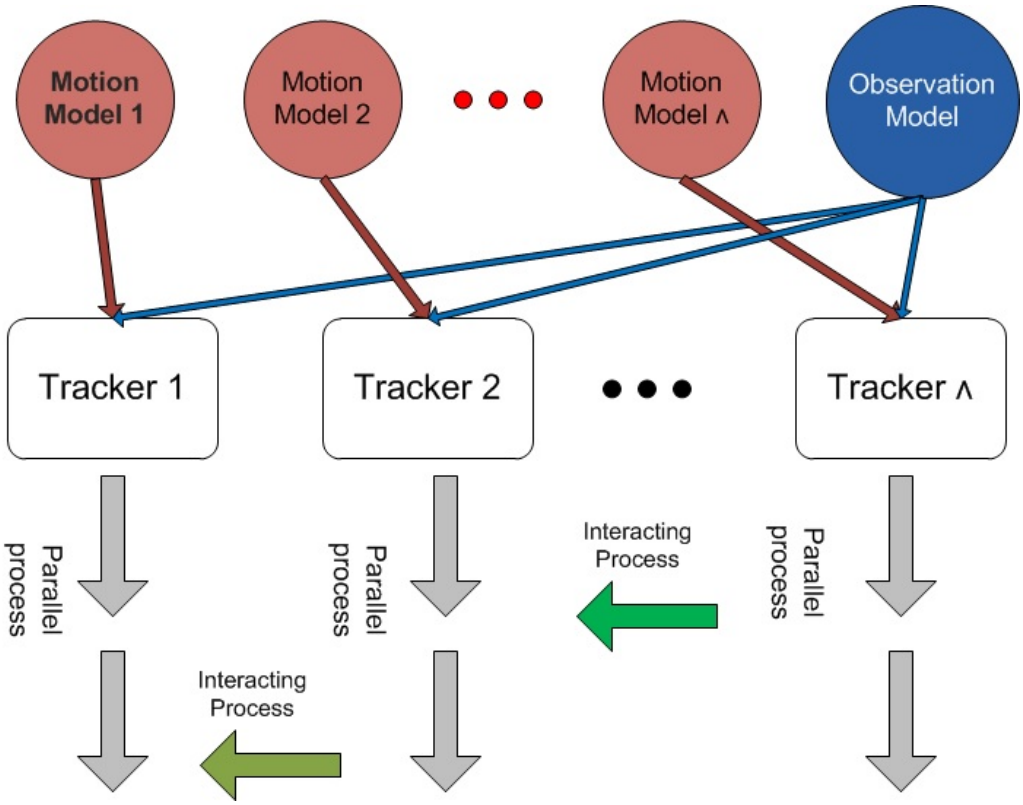


Figure 6.3: Overview of the proposed tracking framework.

6.2 Compound Tracker: Integrating Decomposed Motion Models

Given each decomposed motion model, a basic tracker is composed of a pair of observation and motion models as illustrated in Fig. 6.3. This generates a basic tracker that uses the observation model $p(y_t|x_t)$ and the motion model $p_\lambda(x_t|x_{t-1})$, describing a specific social interaction mode d_λ. The total number of basic trackers for a pedestrian i is the same as the maximum number of motion models Λ. 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 [32]. The Interactive Markov Chain Monte Carlo (IMCMC) algorithm is leveraged to sample across the basic trackers and combine their sampling results. The IMCMC algorithm consists of two main parts: parallel process and interacting process. The Metropolis Hasting algorithm is executed in parallel across all chains during the parallel process while the optimal state value is determined during the interacting process through communication between all chains. The weight w_t^n is implicitly estimated during the interacting process. The implementation used in this dissertation follows the approaches proposed in [39] and Corander *et al.*. [21]. Each Markov chain i runs the MAP estimate given in Eq. 6.6 via the Metropolis Hasting algorithm. The algorithm includes the proposal stage and the acceptance stage. In the proposal stage, a new state x_t^* is proposed by the proposal density function which is Gaussian distribution in our implementation. The acceptance stage decides whether the proposed state is accepted or not with acceptance ratio:

$$\nu_{parallel}(x_t^*|x_t) = min\left(1, \frac{p_i(y_t|x_t^*)}{p_i(y_t|x_t)}\right). \tag{6.11}$$

Here the prior is uniform and hence, it cancels and does not show in the Metropolis Hasting ratio. In interacting process, the Markov chains communicate the others and exchange the states to find better state of an object. A Markov chain accepts the state of another chain β as its state with the following probability:

$$\nu_{interacting} = \frac{p_\beta(y_t|x_t^\beta)}{\sum_{\lambda=1}^{\Lambda} p_\lambda(y_t|x_t^\lambda)}. \tag{6.12}$$

where $p_\beta(y_t|x_t^i)$ and $p_\lambda(y_t|x_t^j)$ return the likelihood scores of Markov chain β and λ.

6.3 Interaction Mode Prediction

Given the set of potential social interaction modes for each pedestrian i, denoted by $\mathbb{D}_i = \{d_1, d_2, \ldots, d_\Lambda\}$, where the size of the set is same across all pedestrians, the interaction model is predicted as summarized in algorithm 2. The number of pedestrians is denoted by H while the number of both social interaction modes and Markov chains is denoted by Λ.

Algorithm 2: Interaction Mode Prediction

Input: Interaction mode sets $\{\mathbb{D}_1, \ldots, \mathbb{D}_i, \ldots, \mathbb{D}_H\}$.
Output: Prediction of Interaction Mode $\{p_1, \ldots, p_i, \ldots, p_H\}$.

```
for Markov chain j = 1 to Λ do
    for Object i = 1 to H do
        predict the new state x_i^j with Equation 6.8;
    end
end
for Iteration number 1 to N do
    Randomly choose one object;
    for Markov chain 1 to Λ do
        Do Monte Carlo sampling using [39];
    end
    Do interaction among all the Markov chains using [39];
end
Estimate the MAP state x̂_1, ..., x̂_i, ..., x̂_H;
Determine the predicted interaction mode p_i based on minimizing the
Euclidean distance between x̂_i and x_i^1, ..., x_i^j, ..., x_i^Λ
```

6.4 Experiments

To evaluate the merit of our proposed model, we perform experiments on both synthesized data and real scenes. Synthesized data is generated to evaluate various parameters in the model. Real scenes are tested to compare the performance of the proposed method against different existing trackers as well as to compare the effect of different functions that could be used to model social forces. Two video sequences were included from the "BEHAVE" Interactions Test Case Scenarios [46]. The videos were acquired at 25 frames/sec and the tracking cycle used is 0.04 s. Two campus pedestrian sequences from the "EPFL" dataset [25] were also included in our analysis. These videos were acquired at 25 frames/sec and the tracking cycle used is 0.04 s. Finally, "QIL" dataset including two video sequences were acquired by our team in an outdoor passageway with six pedestrians in the scene. The videos were acquired at 30 frames/sec and the tracking cycle is 34 ms. The resolution of each frame in the video is 704×480 pixels.

We outline the parameters of our model based on a set of fixed parameters (Table 6.1) and parameters whose values are chosen to optimize tracking performance (Table 6.2). The tracking performance is measured based on CLEAR MOT metrics [13]. We report multiple object tracking precision (MOTP), miss rate (MISS), false positive rate (FP), number of ID switches (IDSW) and multiple object tracking accuracy (MOTA). The threshold for building a matched pair between a tracking result and the ground truth is selected as half of the bounding rectangles' diagonal in the ground truth. It should be noted that MOTP measures the ability of a tracker to

estimate precise pedestrian positions, which is independent of an algorithm's tracking accuracy. MOTP is computed as the average error of the center position of matched pairs' over all frames, measured in pixels. The proposed tracker is named "*Social Interaction-based Multi-target Tracker with three social effects*" (SIMT-3E). We also build one variant of our method which has only two social effects: repulsion and attraction $\hat{\gamma} = \{+, -\}$. This is denoted as SIMT.

Table 6.1: Fixed Parameters of the model

Notation	Meaning	Value	Ref
b	boundary of social force	3 m	
r	radius of pedestrian's private sphere	0.2 m	[43]
m	mass of pedestrian	80 kg	[43]
F^a	magnitude of f^a	500 N	
F^r	magnitude of f^r	500 N	
Δt	tracking cycle	1/(frames per second)	
σ^2	variance of motion model	2	[39]

Table 6.2: Augmented Parameters of the model

Notation	Meaning
$\phi_+(\cdot)$	social force model of attraction effect
$\phi_-(\cdot)$	social force model of repulsion effect
T	size of sliding window

6.4.1 Social mode prediction

The reliability and accuracy of social mode prediction plays a key role in our proposed tracker. We use both synthetic and real scene experiments to tune and assess the prediction performance of our tracker.

6.4.1.1 Synthetic experiment

We consider the social force model [33] as stimulus for synthetic object's motion. Since the stimulus is provided by either the repulsion or attraction force, we employ the SIMT tracker to evaluate social mode prediction. The parameters including magnitude, mass, and boundary and share the same values as our motion model. The synthetic video simulates the object motion video in top view and the corresponding 3D view of the synthetic scene is shown in Fig. 6.4. The videos represent a closed space of 5×5 meters that is walled and has only one opening. In addition, we impose a physical force of 2000 N on the boundary that prevents objects from leaving the field of view. The physical force only affect objects' motion when the distance between boundary and object is less than 25 cm. Further, a physical repulsion force of 1000 N is applied to objects to prevent objects from overlapping when two objects' distance is less than 12.5 cm. The social interaction mode between every two objects is randomly drawn from a uniform distribution. For every 100 frames, a new interaction mode is instantiated for each objects pair. We simulated video sequences with certain number of objects that are randomly initialized in the first frame. The initial velocity is set to 0.5 m/s and the direction of motion is set towards one of four destination points that are located in the middle of four boundaries. In the following frames, objects' motion is computed entirely based on the sum of social and physical forces. The social force is calculated based on the generated interaction mode. Each object is driven by the sum of social interaction force according to Equation 6.8.

To evaluate the prediction accuracy, we generated nine classes of video sequences in which the object number of objects is varied from 2 to 10. Each class includes

five video sequences generated with random initialization. Every video sequence is 5 frames/sec and has 500 frames. The tracking algorithm uses the same configuration shown above. One of synthetic objects' trajectory is shown in Fig. 6.5. On applying our tracking algorithm on the video, the sample of interaction prediction results are shown in Fig. 6.6, which represents the comparison graph between interaction prediction and ground truth. Same color for prediction result and ground truth in one frame indicates a correct prediction, otherwise, it indicates a wrong prediction. We apply our proposed tracker on the synthetic video sequences. The prediction rate and error bar with increasing number of target objects is shown in Fig. 6.7. Note that, the tracking performance stabilizes beyond 5 target objects.

To fully examine the formulation of force model, we test two different force functions on synthetic data, specifically the step function and linear function against the default exponential function:

$$linear(\phi_{-}(i \rightarrow j)) = \left| F^r * (\frac{r_{ij} - d_{ij}}{b}) \right| \mathbf{u}_{ji},$$

$$linear(\phi_{+}(i \rightarrow j)) = \left| F^a * (\frac{d_{ij} - r_{ij}}{b}) \right| \mathbf{u}_{ij},$$

$$step(\phi_{-}(i \rightarrow j)) = \begin{cases} 0 & \text{if } r_{ij} - d_{ij} > b \\ (F^r * \frac{1}{3})\mathbf{u}_{ji} & \text{if } b \geq r_{ij} - d_{ij} > b * \frac{2}{3} \\ (F^r * \frac{2}{3})\mathbf{u}_{ji} & \text{if } b * \frac{2}{3} \geq r_{ij} - d_{ij} > b * \frac{1}{3} \\ F^r \mathbf{u}_{ji} & \text{if } b * \frac{1}{3} \geq r_{ij} - d_{ij} > 0 \end{cases}$$

$$step(\phi_+(i \to j)) = \begin{cases} 0 & \text{if } r_{ij} - d_{ij} > b \\ F^a \mathbf{u}_{ij} & \text{if } b \geq r_{ij} - d_{ij} > b * \frac{2}{3} \\ (F^a * \frac{2}{3})\mathbf{u}_{ij} & \text{if } b * \frac{2}{3} \geq r_{ij} - d_{ij} > b * \frac{1}{3} \\ (F^a * \frac{1}{3})\mathbf{u}_{ij} & \text{if } b * \frac{1}{3} \geq r_{ij} - d_{ij} > 0 \end{cases}$$

where the parameters are same as Eq. 6.3 and Eq. 6.4. Along with different social force functions, we also use a smoothing window and apply the most frequent value filter to reduce the frequency of changes in the predicted interaction mode. Our model predicts the position of a tracked object based on predicting the social mode for each frame in the video. Nonetheless, it is difficult to imagine that people would change their intent, and as a result their motion direction, at each time step. It is hence reasonable to assume that the intent and thereby the predicted social mode under the model would remain unchanged over short time intervals. To enforce this notion, we use the smoothing window of size T, which is measured in the number of frames. Evaluating across all video sequences, the optimal window size are $T = 16$ for exponential function (Fig. 6.8), $T = 12$ for step function, and $T = 3$ for linear function. The comparison of different functions over corresponding optimal window size are shown in Fig. 6.9. From the accuracy curve, we found the force formulation of exponential function results in the most robust prediction performance.

6.4.1.2 Tracker performance in real scene

To validate the result of predication accuracy on real scene, we compare SIMT with three variants that are "SIMT with optimal window" (WSIMT), "SIMT with linear

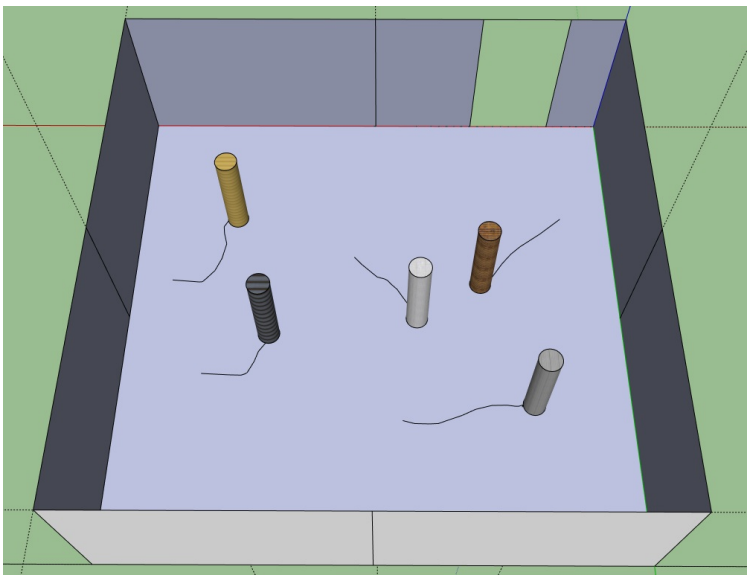


Figure 6.4: 3D view image of synthetic scene.

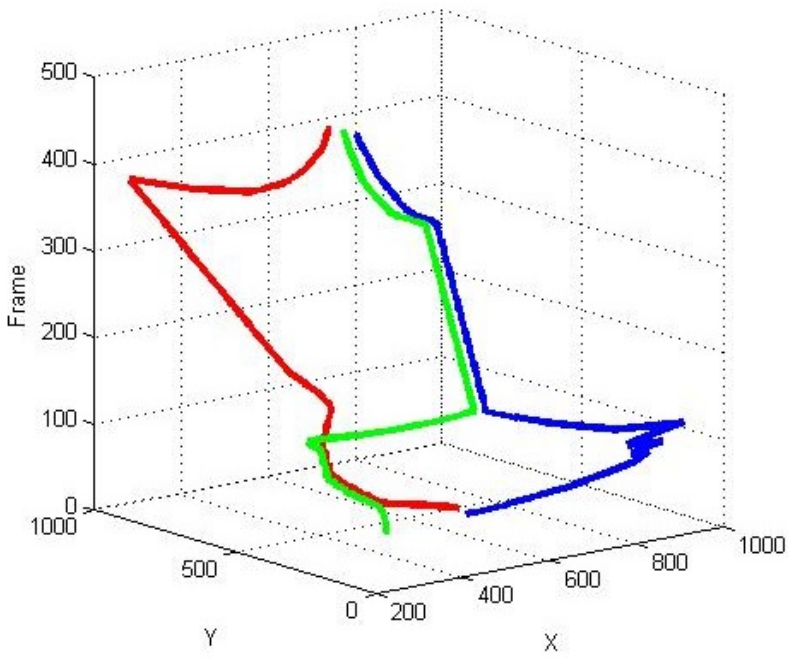


Figure 6.5: Trajectory sample of video sequences with synthetic object interaction. X and Y axes are width and length in image coordinate and the unit is pixels.

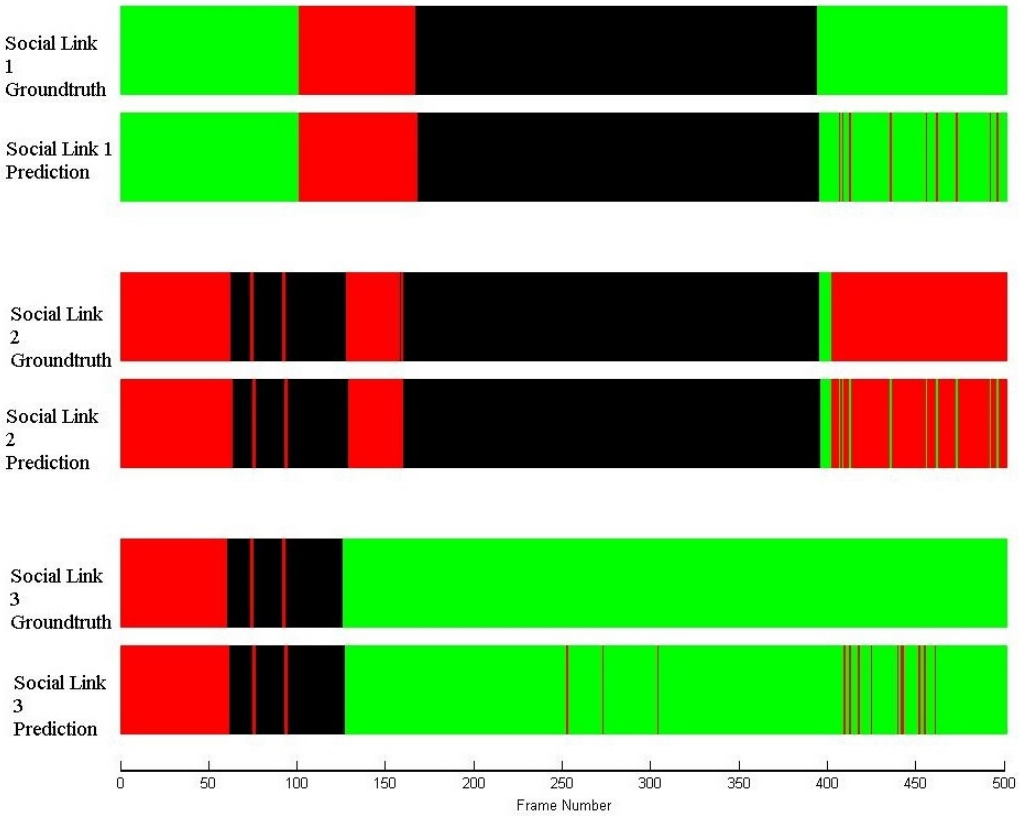


Figure 6.6: Comparison between interaction prediction and ground truth. Red color indicates repulsion effect, green color indicates attraction effect, black indicates there is no interaction.

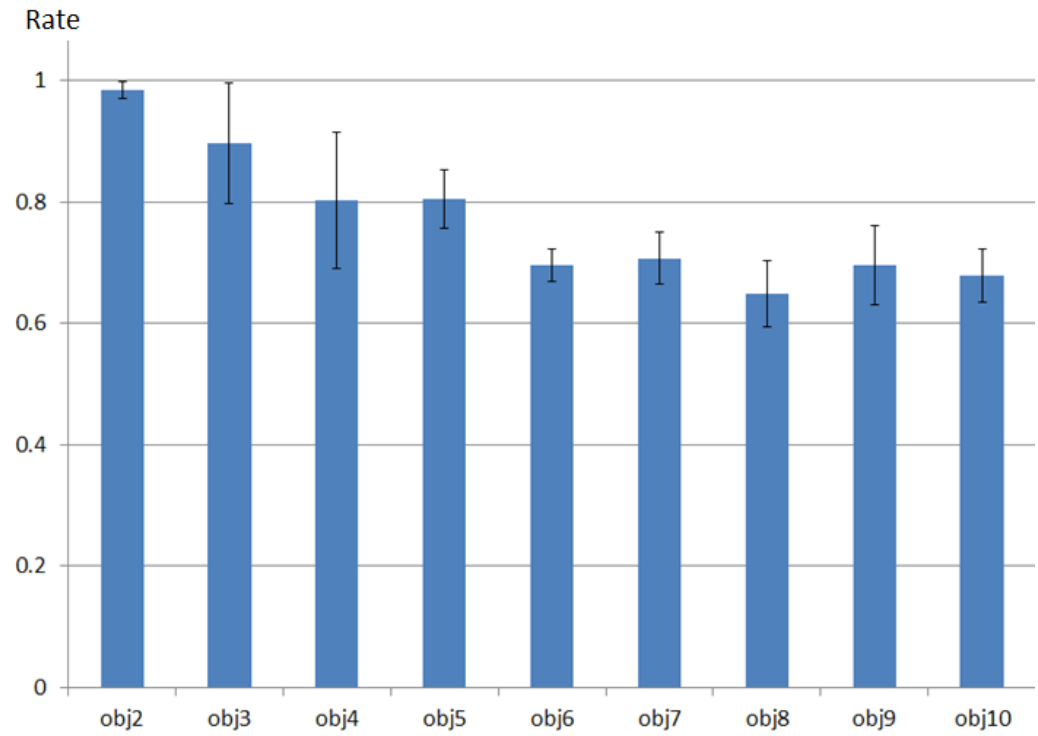


Figure 6.7: Prediction rate of the exponential function with error bar for different classes.

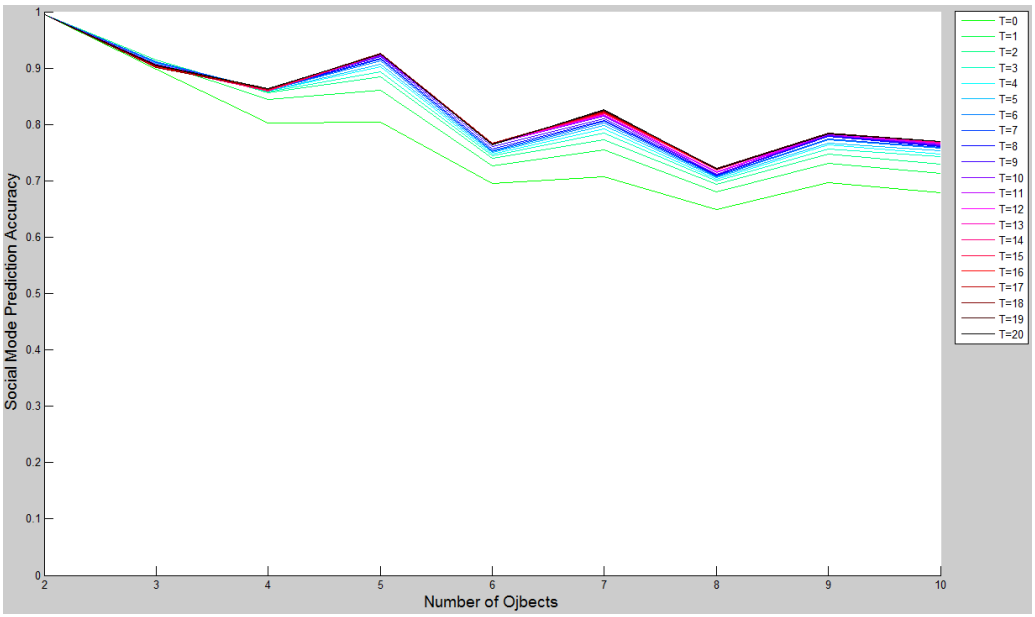


Figure 6.8: Prediction accuracy of the exponential function under different smoothing windows.

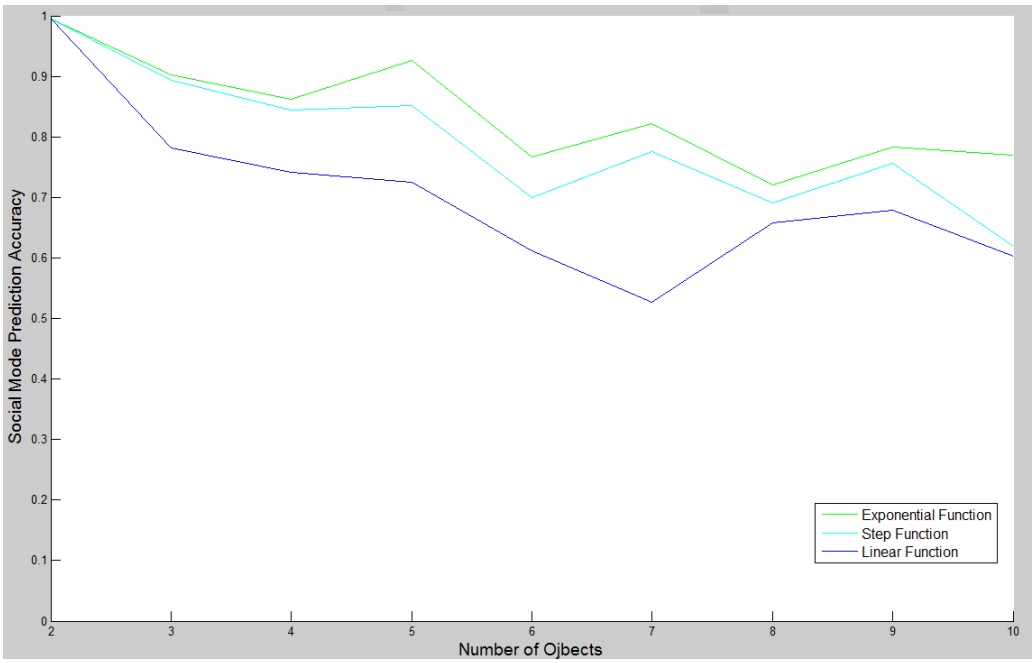


Figure 6.9: Prediction accuracy of optimal smooth window Vs Different social force function.

social force and optimal window" (LnrWSIMT), and "SIMT with step social force and optimal window" (StpWSIMT). Note here that all the trackers use only two social effects ($\hat{\gamma}$=attraction or repulsion). We use a deterministic strategy that chooses the prediction position of smoothed intent instead of MAP estimation results (Eq. 6.6) when the likelihood of MAP results is under a preset threshold. The results on real scenes are shown in Table 6.3 and Table 6.4. StpWSIMT and LnrWSIMT perform worse than SIMT and WSIMT methods, which indicates that intent force in the form of exponential function capture the change of social interaction more accurately. Further, WSIMT outperforms SIMT in terms of MOTP, which means that our smoothed prediction of social interaction mode provides better guidance of human motion and results in better localization in real scenario as shown in Figure 6.10. Overall evaluation on real scenes and synthetic data shows that exponential force function is better over other social force functions and smoothing is better over no smoothing.

Method	MOTP	MISS	FP	IDSW	MOTA
	BEHAVE Seq#1				
SIMT	3.62	**0.00** %	**0.00** %	**0**	**100**%
WSIMT	**2.67**	**0.00** %	**0.00** %	**0**	**100**%
LnrWSIMT	5.54	2.21 %	2.21 %	1	95.25%
StpWSIMT	5.47	1.83 %	1.83 %	1	96.01%
	EPFL Seq#1				
SIMT	2.71	**0.00** %	**0.00** %	**0**	**100**%
WSIMT	2.62	**0.00** %	**0.00** %	**0**	**100**%
LnrWSIMT	2.85	9.86 %	9.86 %	**0**	80.28%
StpWSIMT	**2.26**	10.14 %	10.14 %	**0**	79.72%
	QIL Seq#1				
SIMT	4.67	**0.32** %	**0.32** %	**0**	**99.36**%
WSIMT	4.01	0.95 %	0.95 %	**0**	98.10%
LnrWSIMT	**2.77**	6.08 %	6.08 %	1	87.79%
StpWSIMT	2.89	2.00 %	2.00 %	**0**	96.00%

Table 6.3: CLEAR MOT metrics of SIMT, WSIMT, lnrWSIMT, and StpWSIMT.

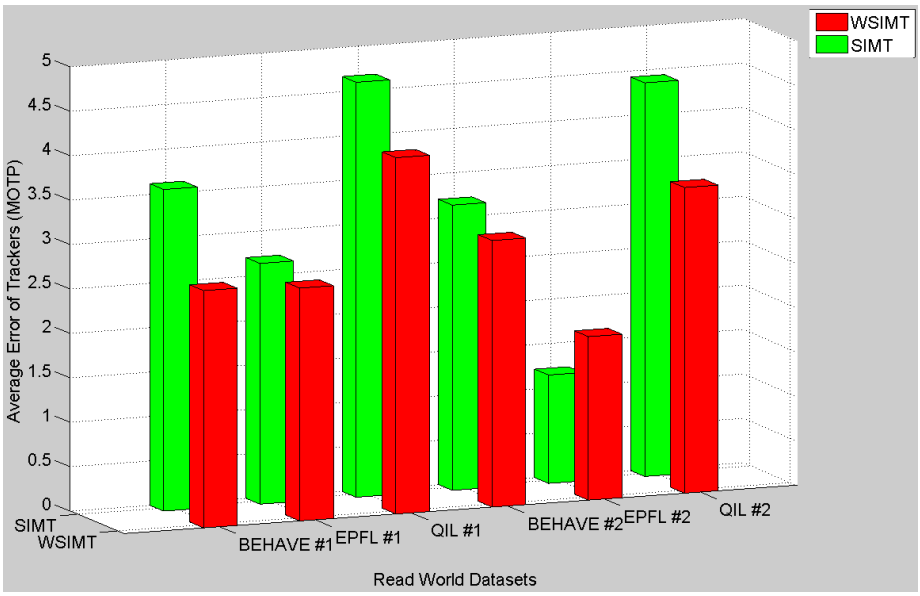


Figure 6.10: WSIMT Vs SIMT.

6.4.2 Social interaction activity analysis

To access the tracking ability and to verify the effectiveness of social effects prediction, we examine our model on common social interaction activity similar as to ones in [56]. We use the our proposed tracker WSIMT-3E, which uses γ (3 social effects) and smoothing window. We create a synthetic experiment dataset, which includes six categories: walking together (Walk), passing (Pass), spreading (Spread), following resulting in walking together (Follow→Walk), following resulting in passing when the followed subject is slower than the other (Follow→Pass), and following resulting in an increasing separation between subject when the followed subject is faster than the other (Follow→Lost). Examples of 4 of these behaviors are shown in Fig. 6.11. The motion speed of synthetic objects is randomly selected from human normal walking speed $0.8 - 1.8\ m/s$. Spreading activity involved $3 - 5$ objects and

Method	MOTP	MISS	FP	IDSW	MOTA
	BEHAVE Seq#2				
SIMT	3.21	**0.00** %	**0.00** %	**0**	**100**%
WSIMT	3.00	**0.00** %	**0.00** %	**0**	**100**%
LnrWSIMT	5.36	8.41 %	8.41 %	**0**	83.18%
StpWSIMT	3.30	**0.00** %	**0.00** %	**0**	**100**%
	EPFL Seq#2				
SIMT	**1.22**	**0.00** %	**0.00** %	**0**	**100**%
WSIMT	1.84	**0.00** %	**0.00** %	**0**	**100**%
LnrWSIMT	2.23	9.37 %	9.37 %	**0**	81.26%
StpWSIMT	2.32	9.37 %	9.37 %	**0**	81.26%
	QIL Seq#2				
SIMT	4.43	**0.00** %	**0.00** %	**0**	**100**%
WSIMT	3.44	**0.00** %	**0.00** %	**0**	**100**%
LnrWSIMT	**3.41**	**0.00** %	**0.00** %	**0**	**100**%
StpWSIMT	3.57	**0.00** %	**0.00** %	**0**	**100**%

Table 6.4: CLEAR MOT metrics of SIMT, WSIMT, lnrWSIMT, and StpWSIMT.

the rest involved two objects. We followed similar scene and object initialization settings as [56] and generated the video sequences automatically. The generated videos are selected to fit our predefined six activity categories and 10 video sequences were synthesized for each activity. To show the effectiveness of social interaction based motion prediction, we replace the motion model Eq. 6.8 with auto regressive motion model [50] while rest of the algorithm was unchanged. The auto regressive motion model represents the human motion as a Gaussian process rather than human interaction. Due to the simple appearance and motion in this synthetic experiment, we evaluated the performance under three sampling settings to highlight the effect of motion model in which the iteration of IMCMC is set to be 50, 250, and 1000, respectively. We run both trackers on synthetic videos and the tracking accuracy and motion prediction precision are measured. The motion prediction precision is defined as the nearest predicted position from groundtruth in the motion models for each objects and the smaller value of precision represents the lower search state space

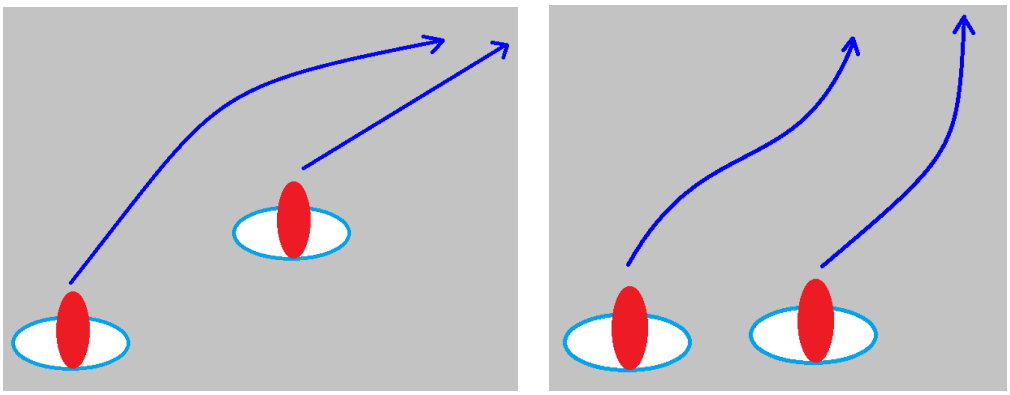
Activities	Iteration 50		Iteration 250		Iteration 1000	
	WSIMT-AR	WSIMT-3E	WSIMT-AR	WSIMT-3E	WSIMT-AR	WSIMT-3E
Walk	4.05 %	4.24 %	97.90 %	97.01 %	100 %	100 %
Pass	66.91 %	72.03 %	100 %	100 %	100 %	100 %
Spread	31.13 %	32.11 %	26.51 %	57.45 %	78.11 %	100 %
Follow→Lost	18.79 %	29.53 %	100 %	100 %	100 %	100 %
Follow→Walk	11.36 %	18.89 %	86.44 %	84.02 %	100 %	100 %
Follow→Pass	6.99 %	9.05 %	75.72 %	76.46 %	100 %	100 %

Table 6.5: Tracking accuracy of synthetic interaction activities. Note that, higher values indicate better accuracy.

Activities	Iteration 50		Iteration 250		Iteration 1000	
	WSIMT-AR	WSIMT-3E	WSIMT-AR	WSIMT-3E	WSIMT-AR	WSIMT-3E
Walk	14.31	13.78	7.05	6.42	6.08	5.07
Pass	3.82	3.56	3.22	3.01	5.07	3.02
Spread	9.60	7.67	8.72	8.50	4.57	3.67
Follow→Lost	9.27	8.14	5.52	4.56	5.82	4.30
Follow→Walk	11.93	9.39	5.57	4.39	5.26	3.98
Follow→Pass	14.42	15.30	6.16	5.39	6.03	4.47

Table 6.6: Prediction precision of synthetic interaction activities. Note that, lower values indicate better precision.

for sampling stage. The comparison of accuracy and precision is provided in Table 6.5 and Table 6.6, respectively. Note here that the precision value only includes the results of correct tracking and is measured in pixel. All the data shown are average value on ten videos sequences of each category. "WSIMT-AR" represents the variant of WSIMT-3E with auto regressive model. The results show the accuracy improves as the iteration number increases and WSIMT-3E is outperform WSIMT-AR in most of the cases. Comparison of precision results clearly shows that WSIMT-3E gives informative prediction of the underlying social effect since the tracker's prediction is closer to the real state of targets, which then reduces the search space in the sampling process.



(a) following (b) walking together

(c) passing (d) spreading

Figure 6.11: Typical scenarios for atomic social interaction activity.

6.4.3 Comparison with various trackers

We evaluated the proposed tracker WSIMT and WSIMT-3E across multiple datasets and compared against several popular visual trackers. Specifically, we compare the tracking results of our method with those of boosted particle filter (BPF) tracker [47], a standard MCMC particle filter tracker [36, 50], and the Visual Tracking Decomposition (VTD) tracker [39]. We also have implemented another tracker which keeps the same tracking framework as ours but replaces the decomposed social interaction model with linear trajectory avoidance model (WSIMT-LTA) [49]. Since the standard MCMC particle filter only uses a single Markov Chain, we perform $N * M$ iterations of the MCMC tracker where N is the maximum number of Markov Chains and M is the number of iterations used in our method. All the trackers are initialized manually by specifying a bounding box in the first frame and data association is entirely based on the generative observation model without any dynamic update.

To initialize the social interaction model, we assume that a pedestrian has a private sphere of radius equal to 0.2 m and a mass of 80 kg. The magnitude of social force is 500 N and the boundary is set to be 3 m. The tracking cycle is equal to the discrete time interval Δt according to every video sequence's frame rate. The variance in Eq. 6.9, σ^2, is set to 2 for all the experiments. The parameters are listed in Table 6.1.

Across all six video sequences, exhibiting varying environments, WSIMT-3E successfully tracked all pedestrians. An illustrative example from one of the "BEHAVE" sequences is shown in Fig. 6.12. This scene shows two pedestrians approaching a

third and eventually forming a group. Four representative frames from each algorithm's tracking results are shown. The proposed tracker and BPF track all pedestrians well before the three pedestrians start exhibiting group interaction. However, only WSIMT-3E continues tracking successfully through the group interaction even though there is significant occlusion. The other two trackers fail due to misleading background and poor image quality. Similar observations can be made for the results from the video sequences in the "EPFL" dataset as Fig. 6.14. Once again the MCMC and the VTD tracker failed early due to complex background and occlusions. BPF tracked well before the occurrence of partial occlusions. In contrast, WSIMT-3E tracks the occluded pedestrians even through abrupt motion changes due to robust prediction based on accountable social interaction modes. Fig. 6.16 shows four representative frames of each algorithm's tracking results from "QIL" dataset. Due to poor image quality, similar appearance among pedestrians, and partial occlusions, MCMC and VTD trackers lose track of several pedestrians. WSIMT-3E and the BPF tracker successfully localize all pedestrians correctly through the video. However, overall, WSIMT-3E exhibits better tracking performance compared to the BPF tracker. Fig. 6.13,6.15,6.17 show the predicted social effects of each social links for the pedestrians with arrow.

To quantitatively compare the result under different scenarios, we manually labeled the ground truth in the six video sequences. Tables 6.7, 6.8, 6.9 present the results of all five algorithms for each of the video sequences from the "BEHAVE", "EPFL", and "QIL", respectively. WSIMT-3E outperforms all the other trackers in

terms of miss rate, false positives rate and ID switches. In terms of MOTP, WSIMT-3E outperforms the other trackers in three video sequences and achieves the second best in three other video sequences. Overall, WSIMT-3E tracks multiple pedestrians more robustly by leveraging individual social interaction modes. The results indicate that the improved motion model also contributes to better localization accuracy. Further, WSIMT-3E improves on precision over WSIMT, which indicates that including the non-interaction social effect within the models leads to better target localization.

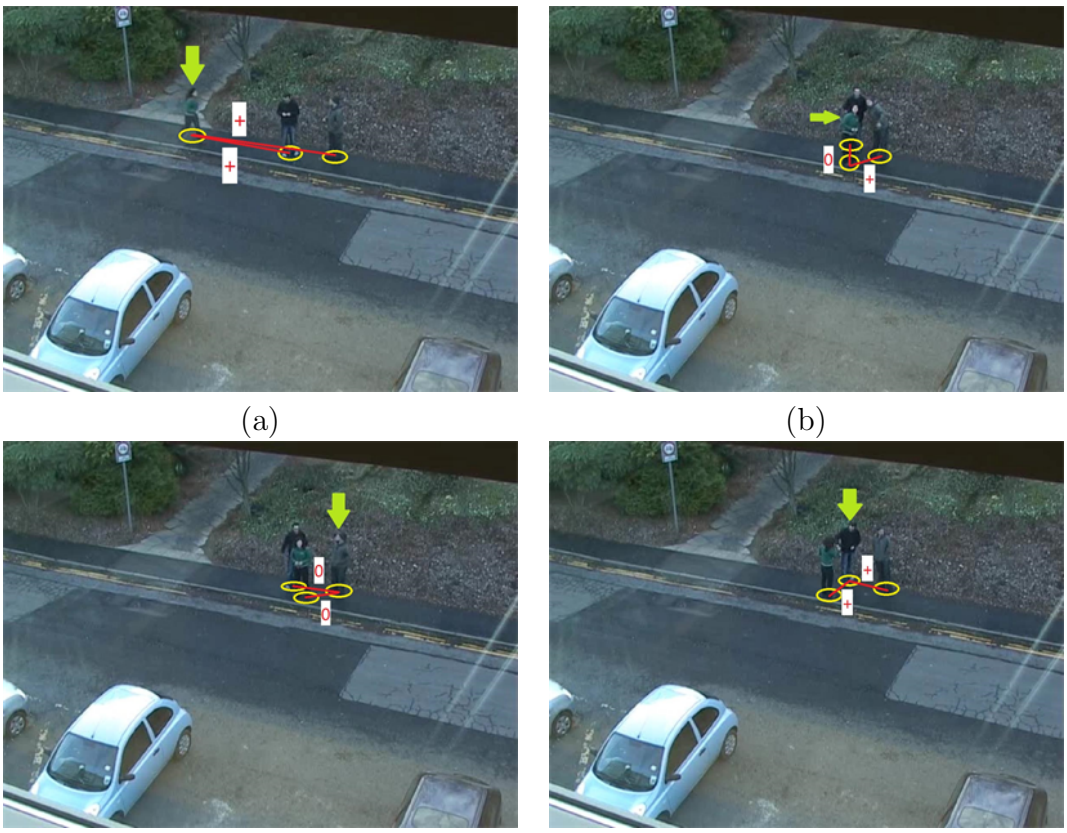
Method	MOTP	MISS	FP	IDSW	MOTA
	Seq#1				
BPF	6.48	54.49 %	54.49 %	**0**	-8.99 %
MCMC	3.86	29.28 %	29.28 %	**0**	41.44 %
VTD	5.46	44.64 %	44.64 %	12	7.54 %
WSIMT-LTA	**2.01**	31.30 %	31.30 %	**0**	37.40 %
WSIMT	2.67	**0.00** %	**0.00** %	**0**	**100** %
WSIMT-3E	2.07	**0.00** %	**0.00**%	**0**	**100** %
	Seq#2				
BPF	3.14	18.21 %	18.21 %	**0**	63.58 %
MCMC	4.99	26.65 %	26.65 %	43	20.33 %
VTD	**2.60**	47.93 %	47.93 %	17	-14.71 %
WSIMT-LTA	3.08	**0.00** %	**0.00** %	**0**	**100** %
WSIMT	3.00	**0.00** %	**0.00** %	**0**	**100** %
WSIMT-3E	2.67	**0.00** %	**0.00** %	**0**	**100** %

Table 6.7: BEHAVE dataset results.

Method	MOTP	MISS	FP	IDSW	MOTA
	Seq#1				
BPF	4.64	8.92 %	8.92 %	**0**	82.16 %
MCMC	3.97	27.48 %	27.48 %	**0**	45.04 %
VTD	4.93	37.12 %	37.12 %	17	24.14 %
WSIMT-LTA	**1.22**	31.44 %	31.44 %	**0**	37.12 %
WSIMT	2.62	**0.00** %	**0.00** %	**0**	**100** %
WSIMT-3E	2.13	**0.00** %	**0.00** %	**0**	**100** %
	Seq#2				
BPF	3.58	9.57 %	9.57 %	**0**	80.86 %
MCMC	14.53	50.43 %	50.43 %	5	-2.32 %
VTD	3.37	38.55 %	38.55 %	2	22.61 %
WSIMT-LTA	5.04	14.04 %	13.16 %	5	71.35 %
WSIMT	1.84	**0.00** %	**0.00** %	**0**	**100** %
WSIMT-3E	**1.71**	**0.00** %	**0.00** %	**0**	**100** %

Table 6.8: EPFL dataset results.

Figure 6.12: The tracking results comparison for selected frames from BEHAVE dataset: BPF (row1), MCMC (row2), VTD (row3), and WSIMT-LTA (row4) WSIMT-3E (row5).



(a) (b)

(c) (d)

Figure 6.13: The predicted social effects of WSIMT-3E for selected frames from BEHAVE dataset.

Figure 6.14: The tracking results comparison for selected frames from EPFL dataset: BPF (row1), MCMC (row2), VTD (row3), and WSIMT-LTA (row4) WSIMT-3E (row5).

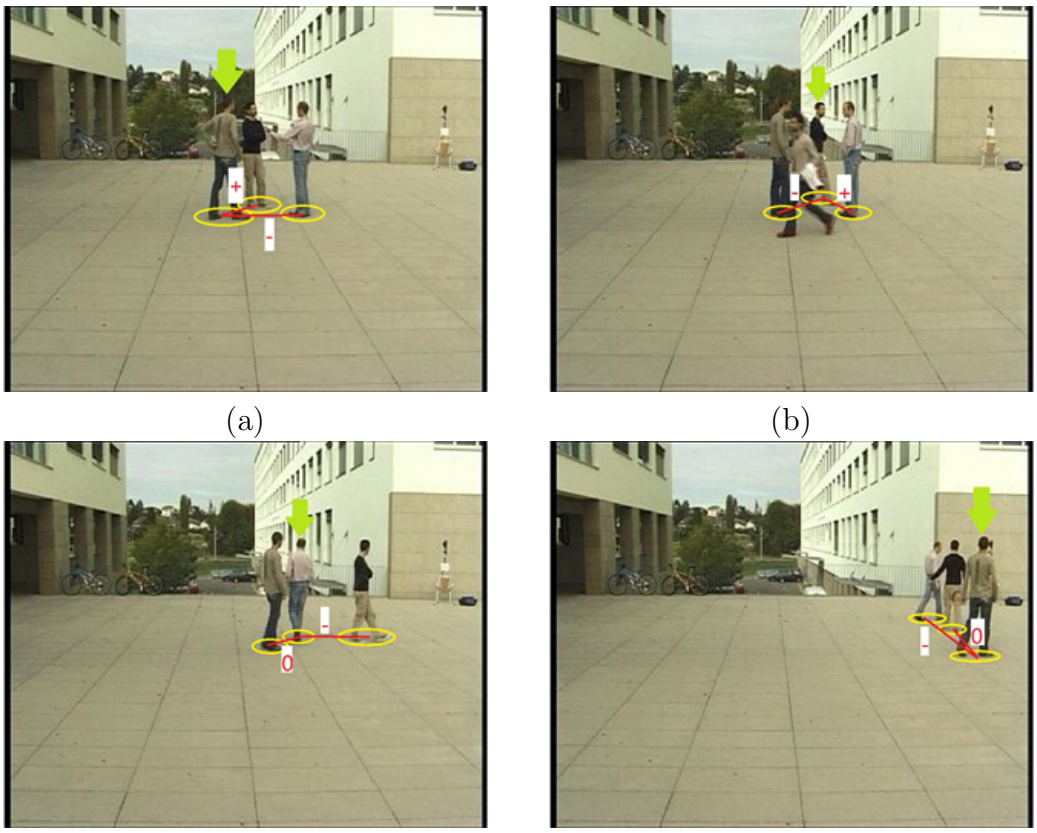
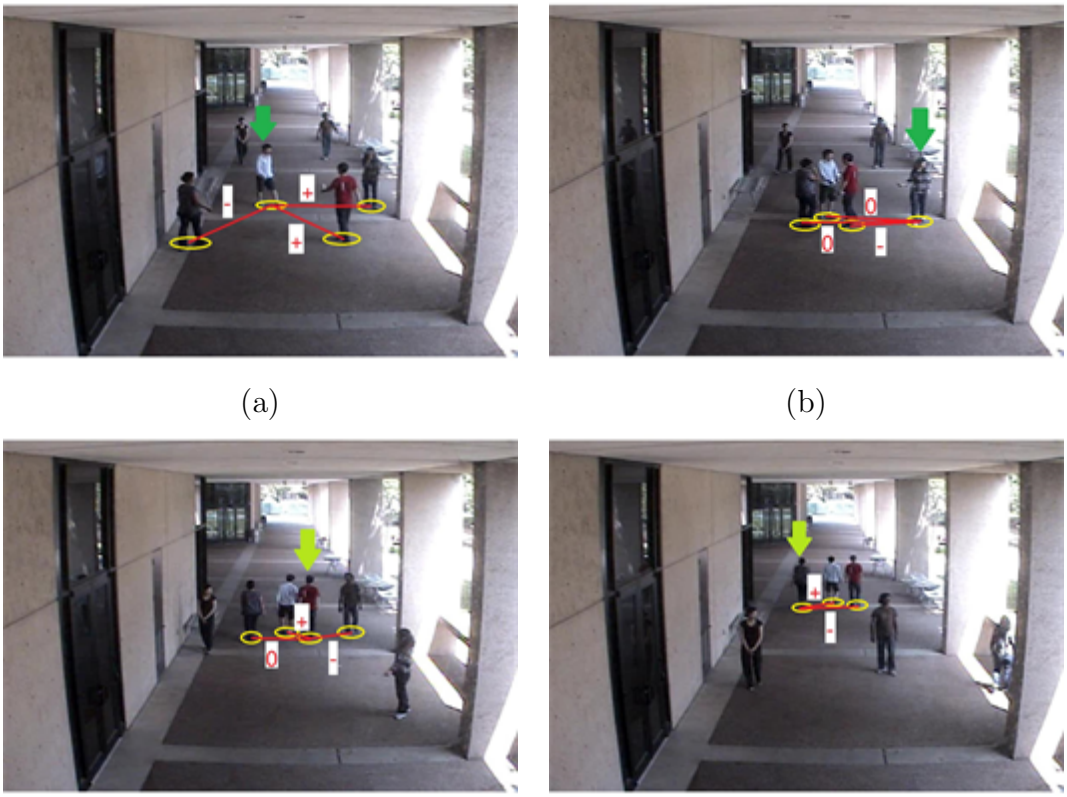


Figure 6.15: The predicted social effects of WSIMT-3E for selected frames from EPFL dataset.

Figure 6.16: The tracking results comparison for selected frames from our dataset: BPF (row1), MCMC (row2), VTD (row3), and WSIMT-LTA (row4) WSIMT-3E (row5).



(a) (b)

(c) (d)

Figure 6.17: The predicted social effects of WSIMT-3E for selected frames from our dataset.

Method	MOTP	MISS	FP	IDSW	MOTA
	Seq#1				
BPF	5.89	1.64 %	1.64 %	1	94.66%
MCMC	12.85	26.39 %	26.39 %	122	39.20%
VTD	8.57	43.01 %	43.01 %	17	10.47%
WSIMT-LTA	7.96	23.69 %	23.69 %	**0**	52.62%
WSIMT	4.01	0.95 %	0.95 %	**0**	98.10%
WSIMT-3E	**3.88**	**0.92** %	**0.92** %	**0**	**98.16** %
	Seq#2				
BPF	6.20	**0.00** %	**0.00** %	**0**	**100**%
MCMC	3.95	22.25 %	22.25 %	**0**	55.50%
VTD	10.18	30.47 %	30.47 %	353	23.88%
WSIMT-LTA	3.51	7.78 %	7.78 %	**0**	84.44%
WSIMT	3.44	**0.00** %	**0.00** %	**0**	**100**%
WSIMT-3E	**3.09**	**0.00** %	**0.00** %	**0**	**100**%

Table 6.9: QIL dataset results.

Chapter 7

Conclusion and Future Work

7.1 Summary of Work

We have proposed four models for multi-person tracking, each of which addresses one of open challenges discussed in Chapter 2. The key contribution are summarized below:

1. *A novel ensemble framework.* The framework leverages the redundancy and diversity between tracking and detection. Association candidates in this integrated model come from independent trackers and object detector. The best candidate is selected based on a score function that integrates classification confidence, appearance affinity, and smoothness constraints imposed using geometry and motion information. Model parameters of the score function are

discriminatively trained. In order to improve the detection confidence in complex scenes, the framework incorporates an additional target classifier that is also trained discriminatively.

2. *A novel visual perception model.* With novel visual perception model, we presented a tracking method using an attentive vision feature where motion analysis is performed in the first-person view. The attentive vision is created from virtually reconstructed scene. A visual attention map is generated based on attentive vision mechanism, including both static and dynamic components. The most feasible path taken by the person is searched and decided from this constructed map. The predicted motion direction is integrated into data-association tracking with color and motion features. The association is solved by a greedy algorithm.

3. *A hierarchical group structure preserving object tracking method.* The method leverages the group structure to identify the relationship among tracked objects. Inter-group and intra-group relationship is modeled as a two-layer graph structure. The proposed structure model is integrated with HOG feature and solved using dynamic programming. The group structure and parameters are initialized and updated continuously using social discovery. The proposed method enables tracking of pedestrians in complex scenes and shows the advantage of the two-layer graph structure in tracking scenarios over single-layer graph structure.

4. *A new social interaction based tracking method.* The method leverages the

social interaction decomposition to approximate a broader set of human interaction behaviors in unconstrained environments. To the best of our knowledge, this is the first time the social force model has been extended to simultaneously model multiple interaction behaviors in human tracking. The proposed dynamic model is decomposed through the construction of multiple basic trackers, each representative of a motion model defined by the specific social interaction mode. An IMCMC framework is used to combine the predictions from the basic trackers to find the best state at each time step.

7.2 Future Work

In our work, we have demonstrated the potential of ensemble of tracking and detection, visual perception modeling, hierarchical group structure and social interaction based tracking to boost multi-person tracking performance. Following are some of the future directions we taken from this work:

1. Use other machine learning techniques to better combine the tracking and detection by estimate the contribute of different feature precisely. The performance of the algorithm could be improved if we enhance the discriminative model for visual matching in the tracker by on-line metric learning.

2. Explore more optimization strategies for data association. The methods from linear and discrete optimization may contribute to better tracking performance compared with greedy search.

3. Explore the ability to learn the number of layers in group structures automatically and find optimal ways to model group graphs. The hierarchical group structure can be extended to more than two layers that could be derived optimally based on the result of human detection or group detection. Further, different ways to construct the graph can be further evaluated to improve the performance.

4. Use data mining of human motion history to get the basic social effects for variant scene. The basic social effects is hypothesized based on heuristic understanding of human interactions. The ability to learn them automatically can be explored for various scenario.

Bibliography

[1] R. Achanta and S. Susstrunk. Saliency detection for content-aware image resizing. In *Proc. International Conference on Image Processing*, pages 1005–1008, Cairo, Egypt, November 2009.

[2] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 798–805, New York, NY, USA, June 2006.

[3] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, December 2006.

[4] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, Ak, USA, June 2008.

[5] A. Andriyenko and K. Schindler. Multi-target tracking by continuous energy minimization. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, June 2011.

[6] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1926–1933, Povidence, RI, USA, June 2012.

[7] G. Antonini, S. Martinez, M. Bierlaire, and J. Thiran. Behavioral priors for detection and tracking of pedestrians in video sequences. *International Journal of Computer Vision*, 69(2):159–180, 2006.

[8] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 983–990, Miami, FL, USA, June 2009.

[9] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1619–1632, August 2011.

[10] L. Bazzani, M. Cristani, and V. Murino. Decentralized particle filter for joint individual-group tracking. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1886–1893, Providence, RI, USA, June 2012.

[11] B. Benfold and I. Reid. Guiding visual surveillance by tracking human attention. In *Proc. British Machine Vision Conference*, pages 1–11, London, UK, September 2009.

[12] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3547–3464, Colorado Springs, CO, USA, June 2011.

[13] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008(1):246309, May 2008.

[14] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Online multi-person tracking-by-detection from a single, uncalibrated camera. *IEEE Transcations on Pattern Analysis and Machine Intelligence*, 33(9):1820–1833, September 2011.

[15] W. Brendel, M. Amer, and S. Todorovic. Multiobject tracking as maximum weight independent set. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1273–1280, Providence, RI, USA, November 2011.

[16] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):500–513, March 2011.

[17] R. Burkard, M. Dell'Amico, and S. Martello. *Assignment Problems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2009.

[18] A. A. Butt and R. T. Collins. Multi-target tracking by lagrangian relaxation to min-cost network flow. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1846–1853, Portland, OR, USA, June 2013.

[19] C. Chen, R. Ugarte, C. Wu, and H. Aghajan. Discovering social interaction in real work environments. In *Proc. International Workshop on Social Behavior Analysis*, pages 933–938, Santa Barbara, CA, USA, March 2011.

[20] W. Choi and S. Savarese. Multiple target tracking in world coordinate with single, minimally calibrated camera. In *Proc. European Conference on Computer Vision*, pages 553–567, Crete, Greece, September 2010.

[21] 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, December 2008.

[22] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, San Diego, CA, USA, June 2005.

[23] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, September 2010.

[24] S. Filipe and L. A. Alexandre. From the human visual system to the computational models of visual attention: a survey. *Artificial Intelligence Review*, 39(1):1–47, 2013.

[25] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multi-camera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282, February 2008.

[26] W. Ge, R. T. Collins, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):1003–1016, May 2012.

[27] E. Goffman. *Behaviour in Public Places, Notes on the Social Organisation of Gatherings*. The Free Press, 1963.

[28] H. Grabner and H. Bischof. On-line boosting and vision. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 260–267, New York, NY, USA, June 2006.

[29] H. Grabner, C. Leistner, and H. Bischof. Semi-supervised on-line boosting for robust tracking. In *Proc. European Conference on Computer Vision*, pages 234–247, Marseille, France, October 2008.

[30] E. T. Hall. *The Hidden Dimension*, volume 6. Doubleday, 1966.

[31] R. Hari and M. V. Kujala. Brain basis of human social interaction: From concepts ot brain imaging. *Physiological Reviews*, 89(2):453–479, April 2009.

[32] W. Hastings. Monte carlo sampling method using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

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

[34] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *Proc. European Conference on Computer Vision*, pages 788–801, Marseille, France, October 2008.

[35] M. Isard and A. Blake. CONDENSATION: Conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[36] Z. Khan, T. Balch, and F. Dellaert. MCMC-based particle filtering for tracking a variable number of interacting targets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1805–1818, November 2005.

[37] C. Koch and S. Ullman. Shifts in slective visual attention: Towards the underlying neural circuitry. *Human Neurbiology*, 4(4):219–227, 1985.

[38] C. Kuo and R. Nevatia. How does person identity recognition help multi-person tracking? In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1217–1224, Colorado Springs, CO, USA, June 2011.

[39] 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, CA, USA, June 2010.

[40] J. Kwon and K. M. Lee. Tracking by sampling trackers. In *Proc. International Conference on Computer Vision*, pages 1195–1202, Barcelona, Spain, November 2011.

[41] B. Leibe, K. Schindler, and L. V. Gool. Coupled detection and trajectory estimation for multi-object tracking. In *Proc.* 11^{th} *IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, Oct. 14-20 2007.

[42] Y. Li, C. Huang, and R. Nevatia. Learning to associate: hybridboosted multi-target tracker for crowded scene. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2953–2960, Miami Beach, FL, USA, June 2009.

[43] M. Luber, J. Stork, G. Tipaldi, and K. Arras. People tracking with human motion prediction from social forces. In *Proc. International Conference on Robotics and Automation*, pages 464–469, Anchorage, AK, USA, May 2010.

[44] Y. Ma and H. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proc. ACM International Conference on Multimedia*, pages 374–381, Berkeley, CA, USA, November 2003.

[45] X. Mei and H. Ling. Robust visual tracking using l_1 minimization. In *Proc. International Conference on Computer Vision*, pages 1436 – 1443, Kyoto, Japan, September 2009.

[46] U. of Edinburgh. Behave interactions test case scenarios, October 2007.

[47] K. Okuma, A. Taleghani, and N. Freitas. A boosted particle filter: multitarget detection and tracking. In *Proc. European Conference on Computer Vision*, pages 28–39, Prague, Czech Republic, May 2004.

[48] N. Ouerhani. *Visual attention: from bio-inspired modeling to real-time implementation*. Ph.D. thesis, Univeristy of Neuchâtel, Switzerland, 2003.

[49] S. Pellegrini, A. Ess, K.Schindler, and L. van Gool. You'll never walk alone: modeling social behavior for multi-target tracking. In *Proc. International Conference on Computer Vision*, pages 261–268, Kyoto, Japan, September 2009.

[50] P. Perez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *Proc. European Conference on Computer Vision*, pages 661–675, Copenhagen, Denmark, May 2002.

[51] M. Piccardi. Background subtraction techniques: a review. In *Proc. IEEE International conference on Systems, Man and Cybernetics*, pages 3099–3104, Manchester, UK, October 2004.

[52] Z. Qin and C. R. Shelton. Improving multi-target tracking via social grouping. In *Proc. IEEE Computer Society on Computer Vision and Pattern Recognition*, pages 1972–1978, Providence, RI, USA, June 2012.

[53] C. Rasmussen and G. D. Hager. Probabilistic data association methods for tracking complex visual objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):560–576, June 2001.

[54] V. Richmond, J. McCroskey, and M. Hickson. *Nonverbal Behavior in Interpersonal Relations*. Pearson/Allyn and Bacon, 2007.

[55] D. A. Ross, J. Lim, R. Lin, and M. Yang. Incremental learning for robust visual tracking. *International Journal of Computer Visiona*, 77(1):125–141, 2008.

[56] K. K. Roudposhti and J. Dias. Probabilistic human interaction understanding: Exploring relationship between human body motion and the environmental context. *Pattern Recognition Letters*, 34(7):820–830, May 2013.

[57] D. Schulz, W. Burgard, D. Fox, and A. Cremers. Tracking multiple moving targets with a mobile robot using particle lters and statistical data association. In *Proc. International Conference on Robotics and Automation*, pages 1665 – 1670, Seoul, Korea, May 2001.

[58] E. L. Schwartz, D. N. Greve, and G. Bonmassar. Space-variant active vision: Definition, overview and examples. *Neural Networks*, 8(7):1297 – 1308, 1995.

[59] G. A. F. Seber. *Multivariate Observations*. John Wiley & Sons, Inc, Hoboken, NJ, 1984.

[60] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated subgradient solver for SVM. In *Proc. International Conference on Machine Learning*, pages 807–814, Corvallis, OR, USA, June 2007.

[61] B. Song, T. Jeng, E. Staudt, and A. K. Roy-Chowdhury. A stochastic graph evolution framework for robust multi-target tracking. In *Proc. European Conference on Computer Vision*, pages 605–619, Heraklion, Crete, Greece, 2010.

[62] S. Stalder, H. Grabner, and L. V. Gool. Cascaded confidence filtering for improved tracking-by-detection. In *Proc. European Conference on Computer Vision*, pages 369–382, Marseille, France, October 2010.

[63] M. Thiebaux, A. Marshall, S. Marsella, and M. Kallman. Smartbody: Behavior realization for embodied conversational agents. In *Proc. International Conference on Autonomous Agents and Multiagent Systems*, pages 1151–1158, Estoril, Portugal, May 2008.

[64] K. Tran, A. Gala, I. Kakadiaris, and S. Shah. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. *Pattern Recognition Letters*, 2013.

[65] V. J. Traver and A. Bernardino. A review of log-polar imaging for visual perception in robotics. *Robotics and Autonomous Systems*, 58(4):378 – 398, April 2010.

[66] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, December 2005.

[67] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: training object detectors with crawled data and crowds. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1449–1456, Providence, RI, USA, June 2011.

[68] X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *Proc. International Conference on Computer Vision*, pages 32–39, Kyoto, Japan, September 2009.

[69] B. Wu and R. Nevatia. Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. *International Journal of Computer Vision*, 82(2):185–204, April 2009.

[70] X. Yan, I. Kakadiaris, and S. Shah. Predicting social interactions for visual tracking. In *Proc. British Machine Vision Conference*, pages 102.1–102.11, Dundee, UK, September 2011.

[71] B. Yang and R. Nevatia. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1918–1925, Providence, RI, USA, June 2012.

[72] B. Yang and R. Nevatia. An online learned CRF model for multi-target tracking. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2034–2041, Providence, RI, USA, June 2012.

[73] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song. Recent advances and trends in visual tracking: A review. *Neurocomputing*, 74(18):3823–3831, November 2011.

[74] A. Yao, D. Uebersax, J. Gall, and L. V. Gool. Tracking people in broadcast sports. In *Proc. 32nd Annual Symposium of the German Association for Pattern Recognition*, pages 151–161, Darmstadt, Germany, September 2010.

[75] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4):13–21, December 2006.

[76] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, AK, USA, June 2008.

[77] L. Zhang and L. van der Matten. Structure preserving object tracking. In *Proc. IEEE Computer Society on Computer Vision and Pattern Recognition*, pages 1838–1845, Portland, OR, USA, June 2013.

[78] T. Zhao and R.Nevatia. Tracking multiple humans in crowded environment. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 406–413, Washington, DC, USA, June 2004.

[79] S. K. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *IEEE Transactions on Image Processing*, 13(11):1491–1506, November 2004.