# CONJOINT INDIVIDUAL AND GROUP TRACKING FRAMEWORK WITH ONLINE LEARNING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

AHMET YİĞİT

IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

FEBRUARY 2016

# CONJOINT INDIVIDUAL AND GROUP TRACKING FRAMEWORK WITH ONLINE LEARNING

Submitted by **AHMET YİĞİT** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Information Systems, Middle East Technical University** by,

Prof. Dr. Nazife Baykal
Director, **Informatics Institute, METU** _____

Prof. Dr. Yasemin Yardımcı Çetin
Head of Department, **Information Systems, METU** _____

Assoc. Prof. Dr. Alptekin Temizel
Supervisor, Modeling and Simulation**, METU** _____

**Examining Committee Members:**

Assoc. Prof. Dr. Altan Koçyiğit
Information Systems, METU _____

Assoc. Prof. Dr. Alptekin Temizel
Modeling and Simulation, METU _____

Assist. Prof. Dr. Aykut Erdem
Computer Engineering, Hacettepe University _____

Assoc. Prof. Dr. Banu Günel
Information Systems, METU _____

Assist. Prof. Dr. Behçet Uğur Töreyin
Informatics Institute, Istanbul Technical University _____

**Date:** 2 February 2016

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

**Name, Last name:** **Ahmet YİĞİT**

**Signature:** _____

# ABSTRACT

## CONJOINT INDIVIDUAL AND GROUP TRACKING FRAMEWORK WITH ONLINE LEARNING

Yiğit, Ahmet

PhD, Department of Information Systems

Supervisor: Assoc. Prof. Dr. Alptekin Temizel

February 2016, 92 pages

A group is a social unit which consists of people interacting with each other and sharing the similar characteristics. Because of social properties of group, group tracking requires taking into account not only visual properties but also social properties such as interaction of people with each other. Also, people groups are dynamic entities and they may grow and shrink with merge and split events. This dynamic nature makes it difficult to track groups using conventional trackers. Besides these difficulties, different types of groups require different strategies in order to perform tracking effectively. While it is possible to track individuals separately when group is sparse, group is considered as a single entity when it is dense.

To overcome and address these challenges, we propose a new tracking strategy, named the Conjoint Individual and Group Tracking (CIGT), based on particle filter and online learning from discriminative appearance model in this thesis. The CIGT proposes a multi observation model with in-group and out-group weights in order to track groups and to evaluate merge and split events. CIGT has two complementary phases: tracking and learning. In the tracking phase, the CIGT calculates multiple weights from observations and models individuals and groups with merge and split events. Particle advection is used in the motion model of CIGT to facilitate tracking of dense groups. In the learning phase, reliable tracklets are first created. Then discriminative appearance model, consisting of shape, color and texture features, is

extracted and used in AdaBoost online learning. State estimate is performed for both individuals and groups by using the discriminative learning model.

Keywords: Particle filter, group tracking, discriminative appearance model, online learning, multi-observation model, particle advection

# ÖZ

## ÇEVİRİMİÇİ ÖGRENİM İLE BİREY VE GRUP TAKİP ALTYAPISI

Yiğit, Ahmet

Doktora, Bilişim Sistemleri

Tez Yöneticisi: Assoc. Prof. Dr. Alptekin Temizel

Şubat 2016, 92 Sayfa

Grup birbiriyle etkileşim kuran ve ortak özellikleri paylaşan insanlardan oluşan sosyal bir birimdir. Sosyal yönünden dolayı, grup takip için görsel özelliklerin yanı sıra insanların birbiriyle olan etkileşimlerinin de dikkate alınması gereklidir. Ayrıca grup dinamik bir birim olup, birleşme ve ayrılma olayları ile büyüyüp küçülebilir. Bu zorluklara ek olarak, etkin bir takip gerçekleştirebilmek farklı grup türleri için farklı stratejiler gerekmektedir. Seyrek gruplarda bireyleri tek tek takip etmek mümkünken, yoğun gruplarda, grup tek bir birim olarak düşünülmektedir.

Bu tez çalışmasında, bu zorlukların üstesinden gelmek ve çözmek için, Birleşik Birey ve Grup Takibi (BBGT) olarak adlandırdığımız parçacık filtre ve ayırt edilebilen görünüm modelinde çevrim içi öğrenme tabanlı yeni bir takip yöntemi öneriyoruz. Bu takip yönteminde grup takibi ve birleşme-ayrılma olaylarını değerlendirebilmek için grup içi ve grup dışı ağırlıkları kullanan çoklu gözlem metodu önerilmektedir. BBGT iki ana aşamadan oluşmaktadır: takip ve öğrenme. Takip aşamasında, BBGT gözlemlerden çoklu ağırlıkları hesaplar ve birleşme ve ayrılma olaylarını değerlendirir. Kalabalık grupları takip için hareket modelinde parçacık adveksiyon metodunu kullanır. Öğrenme aşamasında, güvenilir izler önce oluşturulur. Sonrasında ayırt edilen görünüm modelindeki şekil, renk ve doku özel-

likleri çıkarılır ve AdaBoost çevrim içi öğrenmede kullanılır. Ayırt edilen görünüm modeli kullanılarak hem bireyler hem de gruplar için durum tahmini yapılır.

Anahtar Kelimeler: Parçacık filtresi, grup takibi, ayırt edilen görünüm modeli, çevirim içi öğrenme, çoklu gözlem modeli, parçacık adveksiyonu

*Dedicated to My Family*

## ACKNOWLEDGMENTS

During my education life, I have always faced different challenges and struggled to overcome them. Fortunately, I have found great people who helped me without whom I could never finish my education. This rule of thumb was also valid during my PhD journey and I could not finish this work without them.

I would like to express my deepest gratitude to my supervisor Assoc. Prof. Dr. Alptekin Temizel for his guidance, suggestions support and encouragement throughout the whole process. He provided me with great vision and advice to complete my dissertation. He is always a great model for me with his thoughts to my life.

Thanks to my thesis committee; Assoc. Prof. Banu Günel and Assoc. Prof. Dr. Behçet Uğur Töreyin for their valuable comments and suggestions throughout the last few years. Special thanks to Cihan Ongun who helped me on Particle Advection. Also, I would like to thank to Loris Bazzani for helping me on generation of person detection from the ground truth data.

My sincere thanks go to the other thesis jury members Assoc. Prof. Dr. Altan Koçyiğit and Assist. Prof. Dr. Aykut Erdem for reading and commenting on the final version of this thesis.

Last but not least, I am very lucky to have a wonderful family that unconditionally supports me in pursuing my goals. They have always encouraged me even at my difficult times. I want to dedicate my thesis to my family.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| CIGT | : | Conjoint Individual and Group Tracker |
| DPF | : | Decentralized Particle Filter |
| MOTA | : | Multi Object Tracking Accuracy |
| MOTP | : | Multi Object Tracking Precision |
| RSME | : | Rooted Mean Square Error |
| CRF | : | Conditional Random Fields |
| MAP | : | Maximum A Posteriori |
| HOG | : | Histogram of Oriented Gradient |
| SVN | : | Support Vector Machine |
| MST | : | Minimum Spanning Tree |
| SFM | : | Social Force Model |
| LBP | : | Local Binary Pattern |
| MCMC | : | Markov Chain Monte Carlo |
| MHT | : | Multiple Hypothesis Tracker |
| PDA | : | Probability Density Association |
| DPMM | : | Dirichlet Process Mixture Models |
| FTLE | : | Finite-Time Lyapunov Exponents |
| GPU | : | Graphical Processing Unit |
| HMM | : | Hidden Markov Model |
| FP | : | False Positive |
| FN | : | False Negative |
| GLCM | : | Gray Level Cooccurance Matrix |
| RLM | : | Run Length Matrix |
| AR | : | Autoregressive |
| LBP | : | Local Binary Pattern |
| FFT | : | Fast Fourier Transform |

# CHAPTER 1

# INTRODUCTION

## 1.1 Background and Motivation

People tracking plays an important role in video surveillance and is used in many areas such as employee safety, event surveillance, vandalism deterrence, and public safety. Also, in the last decades, it is commonly used for crowd analysis. Especially anomaly detection in airports, subways and malls is very crucial for the security of people. There are many methods to detect abnormal or threatening events automatically. Commonly used methods for crowd analysis are identified as People Tracking, People Counting and Behavior Understanding as shown in Figure 1 [1].

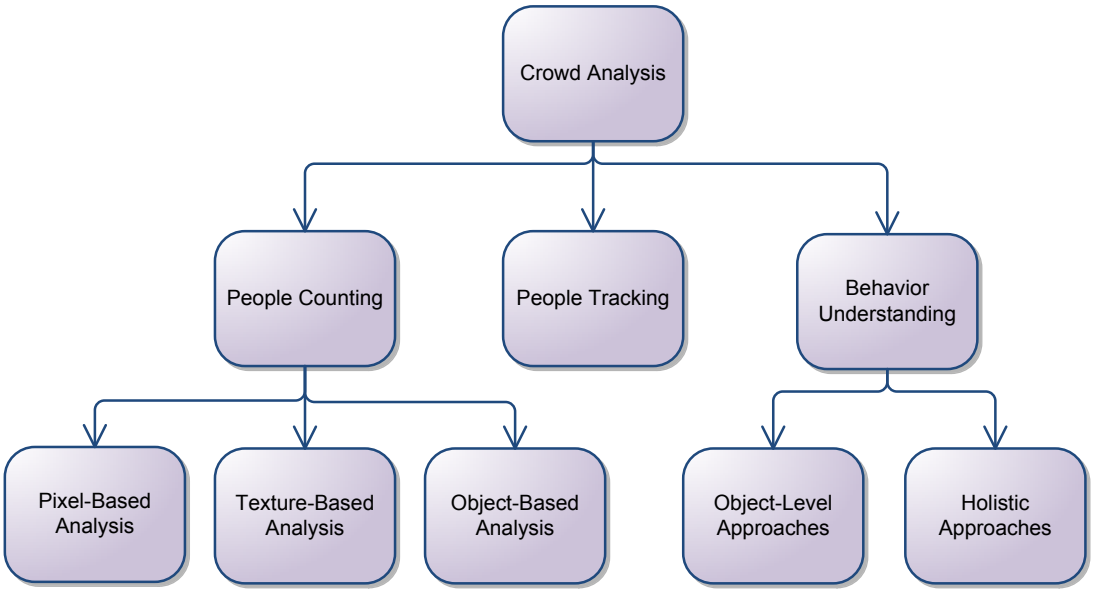


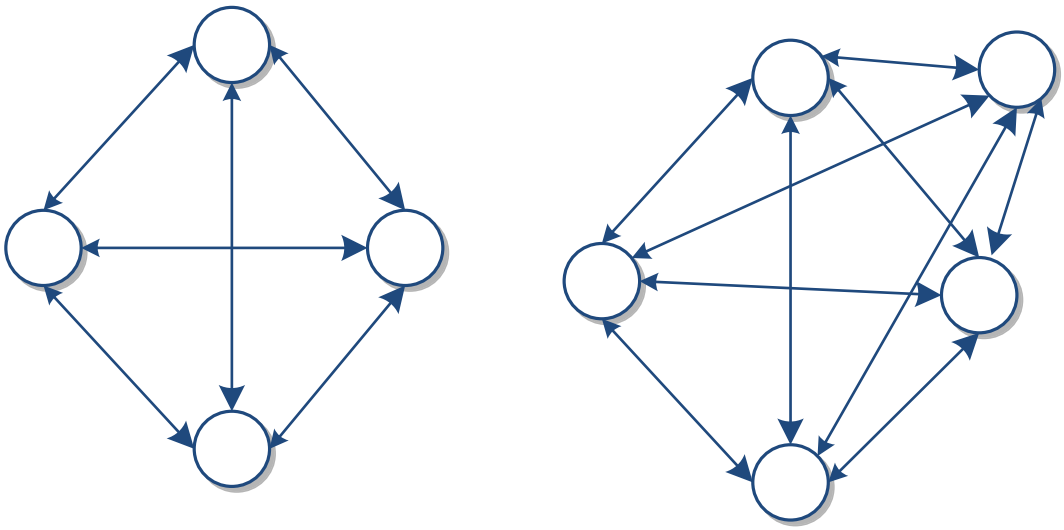

**Figure 1: Crowd Analysis methods according to [1]**

In the context of crowd analysis, behavior understanding is a high level analysis while people tracking and counting are low-level analyses. People tracking and counting are used to extract information used to understand the behavior of crowds. Therefore, people tracking and counting are fundamental parts for higher level behavior understanding.

In this thesis, we mainly focus on the people tracking field and aim to model a tracker for both individuals and groups. In social sciences, a group is defined as a social unit which consists of people who interact with each other and share similar characteristics. Because of the social properties of groups, group tracking requires taking into account not only visual properties but also social properties such as interaction of people with each other. Interactions between people can be modelled with motion similarity and degree of closeness and computer vision algorithms take advantage of these properties to analyze the behavior of a group.

Moreover, groups are dynamic entities and may grow or shrink with merge and split events. This makes it difficult to track groups since the properties of the group used by the tracker may change in time causing tracker drifts or even fails. This problem can be addressed by evaluating the interaction between people and grow or shrink tracked group as a result of the interaction.

Group size is another factor that affects the group structure and interactions. While group size increases, the group becomes more structured and is considered as a single unit. This affects the group motion pattern and tracking algorithms need to evaluate groups in this perspective. Also, group size affects the interactions between people. While group size increases, interaction possibility between the people increases gradually. One way to analyze the interaction possibilities is social interaction graph shown in Figure 2.



(a) Interactions between four people (b) Interactions between five people

**Figure 2: Social Interaction Graph example between people. Nodes represents individuals and edges represent interactions between people**

In Figure 2 (a), while there are six different interactions possibility between four people, the number of interaction possibilities increases to ten when five people exist in the group. In order to evaluate the grouping events in group formation, tracking algorithm also needs to evaluate the effect possibilities as well.

In this thesis, we propose a particle filter based conjoint tracker [2] with multi-observation model for tracking of multiple individuals and groups. We consider an individual as a one-person group and propose that we can track individuals with the same method that we developed to track groups. The proposed multi observation method is inspired from the sociological definition of group in order to model both individuals and groups.

## 1.2 Contributions of Research

In this thesis, our motivation is to develop a unified individual and group tracker that can be used in crowd analysis. The suggested framework assumes that an individual is a one-person group and proposes a tracking framework which works on both groups and individuals.

The major contributions of this study are as follows:

(i) Integrates a multi-observation model which uses in-group and out-group weights in the observation model. This allows identification of different features in a group and allows detection of merge and split events and group formations;

(ii) Particle advection is used to calculate the motion flow. Particle advection method has been shown to be an effective method for analyzing crowd dynamics especially when the crowd density is high and tracking of individuals is not feasible due to occlusions [3, 4].

(iii) Unlike the standard particle filter based method that uses a fixed number of particles during tracking, dynamic particle sampling with respect to group density is used.

(iv) Discriminative Appearance Model [5] is used to identify individuals and is embedded into group state estimate.

(v) Two-phase association is proposed to handle partial occlusion and reduce the number of id switches.

(vi) Hierarchical false positive elimination mechanism is used to reduce false positive detection and increase tracking performance.

The assumptions and limitations of this work are stated as follows:

(i) In this thesis, we focus on self-organizing groups where group members cooperate and interact with each other around some task of interest [6].

(ii) We assume that videos are captured by stationary cameras.

(iii) Although our framework can support more than 40 people group, we limit the group size as 40 due to increasing frame processing time

The remainder of the thesis is organized as follows. In Chapter 2, we provide the background for Particle Filter in general and the related works. Chapter 3 presents Sociological Background of the work. In Chapter 4, we present the proposed method. Chapter 5 presents the experiment results and evaluation, and finally Chapter 6 concludes the thesis.

# CHAPTER 2

# BACKGROUND AND RELATED WORK

Visual tracking is the one of the most important topics used in many surveillance scenarios. Although visual tracking is used in many different areas such as traffic analysis, parking control etc., we specifically focus on people tracking. Therefore, we provide the literature survey in this context.

In this chapter, we provide a brief overview of state-of-the art people tracking methods. First, we provide an overview of the standard particle filter model and describe state model factorization for common particle filter based models and CIGT framework. Then, related works on tracking is explained and examined in detail. Since one of our contributions is the use of particle advection in group tracking, we also provide related works about it. In the final part, we summarize this section.

## 2.1 Particle Filter Overview and State Model Factorization for Tracking

It is very important to provide non-linearity and non-Gaussianity for tracker's accuracy. Particle filter addresses such a system. It uses sequential Monte Carlo methods to estimate the state of the system:

$$x_{t+1} = f_t(x_t, n_t), \qquad y_t = g_t(x_t, m_t) \tag{2.1}$$

where $x_t$ is the state of the system, $y_t$ is the observation, $n_t$ and $m_t$ are non-Gaussian noises, and $f_t$ and $g_t$ are non-linear functions. Since the particle filter uses Markov model, it assumes that observations are dependent only on the current state. By using Markov model and Bayes theorem, it finds the posterior distribution as follows:

$$P(x_t|y_{0:t}) = P(y_t|x_t)\pi(x_t|x_{t-1}) \quad (2.2)$$

where $P(y_t|x_t)$ is the observation model, and $\pi(x_t|x_{t-1})$ is the motion model. Particle filter approximates the filtered posterior distribution by using weighted particles to estimate the system state. Observation model is used to calculate the weight of the particles and motion model is used to move particles.

The state models and their factorizations are very important for tracking and they aim to predict the next state of the tracked object by using prior knowledge. Figure 3 shows the standard particle filter factorization.

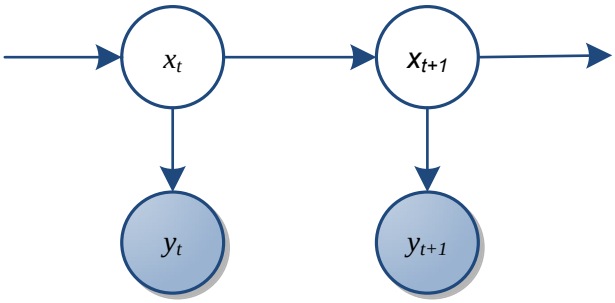



**Figure 3: Standard Particle Filter Factorization**

In Figure 3, $x_t$ is the state of the system and $y_t$ is the observation. Standard particle filter estimates the object state by using the previous state and the current observation. It uses only one measurement as an observation, for which visual similarity is generally used.

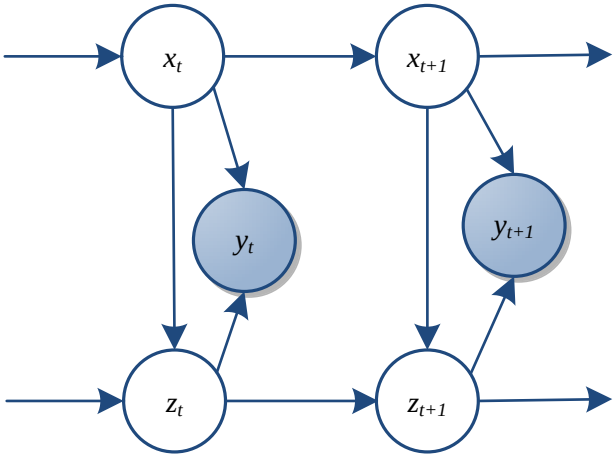



**Figure 4: DEEPER-JIGT Factorization [7]**

In Figure 4, $x_t$ is the individual state, $z_t$ is the group state, $y_t$ is the observation. DEEPER-JIGT [7] is the Decentralized Particle Filter based joint individual and group tracking framework and its factorization [7] decomposes the object state into individual state and group state as shown in Figure 4. Similar to standard particle filter, individual state is estimated by using the previous state and the current observation while group state is estimated from the previous group state, current individual state and current observation measurement. That is, individual state is used in the estimation of the group state.

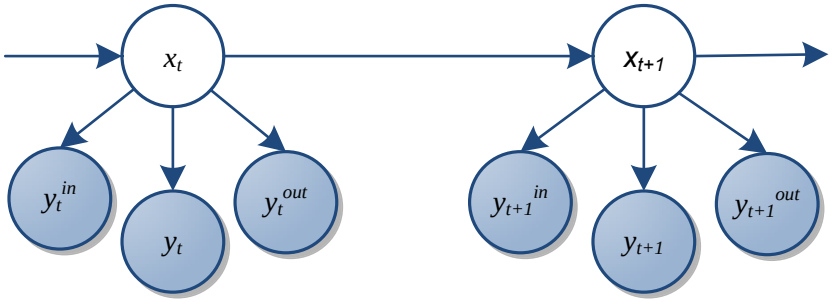



**Figure 5: CIGT Factorization [2]**

In Figure 5, $x_t$ is the system state, $y_t$ is the similarity observation, $y_t^{in}$ and $y_t^{out}$ is in-group and out-group observations, respectively. Unlike DEEPER-JIGT factorization [7], The CIGT Factorization [2], shown in Figure 5, decomposes the observation model instead of the object state. This feature provides CIGT to keep single state information for both groups and individuals. Individuals' interaction is evaluated with in-group and out-group measurements which are used in group evaluation.

## 2.2 Related Works on Tracking Methods

Since our main motivation is to facilitate tracking in scenarios where there are both individuals and groups, we evaluate the tracking methods in the context which can be classified into three main parts: Individual Tracking, Group Tracking, and Individual and Group Tracking methods.

### 2.2.1 Individual Tracking

Individual tracking methods are generally preferred in sparse environments. In the last decades, a number of researches have been exploited to address the pedestrian tracking in dense environments [8, 9] and handling the occlusion problems [5, 8, 10]. In order to address these challenges, different visual properties are used.

In order to solve the occlusion problem, [5] proposes online learning of discriminative appearance model for multi-target tracking in crowded scenes. In this method, an object is represented by means of its texture, shape and color. First, reliable tracklets are created by using [11]. Then, it extracts positive and negative samples by using spatial-temporal constraints, construct weak hypotheses from these samples and build AdaBoost online learning model for appearance representation. Association between detection and tracklets is performed with this learning model.

Different from [5], [12] uses discriminative appearance model on different body parts. In [12], unoccluded parts are found and these parts are removed from the appearance model. The appearance model for each tracklet in [12] consists of the feature set and a weight set for corresponding feature. By using these features and weights, [12] creates AdaBoost online learning model for appearance and solve occlusion problem.

[13] proposes a complete individual tracking framework based on the methodology proposed in [5]. The method consists of three main parts: visual tracking, track management and online learning. In the visual tracking part, it combines the data association model with the particle filtering model. Track existence probability is estimated by data association while the track states are estimated with associated detection using particle filtering. Track management performs track link, termination and initialization by using track existence probability and track state estimates. In online learning step, it uses AdaBoost online learning model and track affinity score is calculated with discriminative appearance, shape (only height and width of object) and motion models.

Besides the appearance model based online learning methodologies in [5, 12, 13], there are trajectory based approaches [14, 15, 16] to solve the occlusion problem. In [14], each target is assigned to a unique trajectory which is the best match to the target's motion. In order to accomplish this, it aims to design energy function. However, this function should reflect all true situation as accurately as possible. Therefore, it builds energy function which consists of five terms: an observation term, collision term, avoidance term, object persistence term and regularizer to keep the number of trajectories low. HOG feature is used in the observation term. Conjugate gradient method is employed to minimize this energy function.

Similar to [14], [15, 16] propose a trajectory based method and address the following problems:

(i) Assigning observations to the correct target;

(ii) Finding and correcting the trajectories of all targets.

In [15], data association is performed with discrete optimization with label cost while trajectory is estimated by solving continuous fitting problem. [16] proposes discrete-continuous Conditional Random Fields (CRF) model in order to solve the occlusion problem in individual tracking. CRF model consists of unary and pairwise terms. In [16], unary term is used to perform the exclusion between trajectories while pairwise term is used in exclusion between observations. These two terms form the energy function. Instead of minimizing this energy function, [16] proposes a MAP based scheme to fit best match.

CRF model is widely used in multi-target tracking domain and an expandable model with appearance model, motion model. [17] also uses the discriminative appearance model and combined with the Conditional Random Fields (CRF) based online learning model. In CRF model, energy function is defined with unary and pairwise functions. In [17], unary functions are used to discriminate all targets while pairwise functions aim to discriminate corresponding tracklet pairs. The occlusion problem is handled in pairwise terms. Both unary and pairwise terms consists of not only the appearance model but also the motion model. Once the energy function is defined with unary and pairwise terms, multi-target tracking problem turns out to energy minimization problem. [17] solves this energy minimization problem by using Hungarian algorithm.

The motion modeling is particularly important for individual tracking in order to handle the occlusion problem. Most of the methods assume constant velocity for the object and use a linear motion model [17]. In addition to the linear motion model, there are other methods in the state-of-the art to model motion. [18] proposes an online method in order to learn a non-linear motion pattern and combines it with the discriminative appearance model. In [18], non-linear motion map is built to better model direction changes and calculate motion affinities between tracklets. A multiple instance learning method is also used for the appearance model for tracking. Unlike other methods, [19] combines people tracker with individual dynamics by using social force model. [19] uses sociological facts to describe the individual motion model, such as the intention towards a goal and constraints from the environment and takes an account of the interaction between people. However, there is no learning mechanism for the appearance model.

We aim to examine the Particle filter based approaches in the context of this thesis. [20] integrates tracking-by-detection approach with particle filtering framework where a continuous confidence is used for the person detector and online learned classifier is used for each indi-

vidual. In this way, it aims to combine the general information of detection with instance-specific information and therefore, builds an observation model as follows:

$$p(y_t|x_t^i) = \underbrace{\beta \cdot I(tr) \cdot p_N(p - d^*)}_{detection} + \underbrace{\gamma \cdot d_c(p) \cdot p_o(tr)}_{\substack{detection \\ confidence\ density}} + \underbrace{\eta \cdot c_{tr}(p)}_{classifier} \tag{2.3}$$

The *detection* term is computed between particle $p$ and the associated detection $d^*$. In the *detection confidence density* term, confidence density $d_c(p)$ at particle $p$ is computed by intermediate object detector used in the observation model. To complete with instance specific information, classifier trained for target $tr$ is used in the observation model.

In [21], a particle filter based framework is combined with the human detector with HOG descriptor in order to build an effective individual tracker. First, the detection responses are associated with the tracker output by matching the color histogram. Then, similar to [20], this associated result and output of HOG based SVN classifier is fused in the observation model.

Particle filter based tracking method is used for not only colored videos but also other types of camera captured videos such as thermal or infrared since the observation model gives the flexibility to use different measurements as an observation weight. [22] proposes a pedestrian tracking method for infrared videos. Different from the color histogram and HOG feature used in [20, 21], intensity and edge cues are integrated as a weighted sum in the observation model.

The other common challenge to be addressed in the state-of-art is the individual tracking in the crowded scene. [8] aims to associate head detections to a set of head tracks in order to perform tracking in crowded scenes. First, it builds an energy function which combines the crowd density estimation with person detection. Then, the energy function is minimized to detect and track individuals by jointly optimizing the density and location of individuals.

Besides evaluating frame-by-frame approaches, space-time constraints and analysis are used for individual tracking approaches to address different challenges. [23] proposes a Bayesian framework for tracking individuals in crowded scenes by using space-time knowledge of the crowd. It trains the Hidden Markov models with spatio-temporal motion patterns obtained from crowd videos which includes different crowd patterns. Then, it predicts the local spatio-temporal motion patterns for individuals tracking. [24] proposes an adaptive appearance model by using spatio-temporal cues. Appearance model consists of several temporal cues

and these temporal cues consist of spatial cues. The proposed method combines this appearance model with particle filtering framework in order to handle the occlusion problem.

In addition to particle filter, there are also different tracking methods used in literature. [19, 25] use the Kalman filter as a tracking engine. [25] combines person detection system based on multi-scale deformable part models [26] with Kalman filtering while [19] uses Kalman filter tracking for social interaction evaluation and individual tracking. [10, 27] use Mean-shift tracker to solve the occlusion problem.

### 2.2.2 Group Tracking

Group tracking is useful when detection and tracking of individuals are not feasible [28] and it aims to model the interactions between people by using different features. Individual tracking can be used as a helper in order to evaluate interactions and group formations. [29] proposes a framework which consists of two particle filter trackers: one for group tracking and one for individual tracking. The group tracker handles groups as atomic entities. The individual tracker is used as a helper to the group tracker and they work collaboratively. In [29], two trackers share the same observation but evaluate this observation differently.

[30] proposes a tracking framework in which multiple pedestrian tracker reflects group behavior of pedestrians by using minimum spanning trees (MST). Firstly, hierarchical clustering is performed on pedestrians in order to form groups. In clustering process of [30], velocity and position of individuals are used as a feature and MST is built according to these features for each group. Then, each group is tracked separately.

Since group is a social unit in which individuals interact with each other, human behavior analysis can be used in group tracking. [31] extends the SFM for group detection and embeds human behavior analysis to predict interaction. Group is formed according to the velocity of group members. By using this approach, it evaluates the interaction detection, group crossing and approaching events in the group formation. Once group is identified, [31] employs a buffered greedy graph-based multi-target tracker [32] in order to track groups.

In addition to [31], [33] proposes a group detection method based on learned social relationship between individuals and forms the group according to this relationship. First, it performs person detection and tracks detected people by using multiple hypothesis tracker (MHT). In group detection, [33] constructs the social network graph between people and edge of graph connects the socially related people via coherent motion indication which con-

sist of relative distance, velocity and orientation differences. In group tracking, it employs multi-model MHT which is extended by the intermediate tree level in group formation. Also, pedestrian detection and tracking methods are used to assist group tracking.

Group tracking methods are used in crowd analysis and behavior understanding methods. [34] propose a method to recognize group behavior and identify violent behaviors in subway scenarios. The method in [34] consists of three main parts: people detection, detection and tracking groups, event detection. In [34], people are detected with AdaBoost trained with LBP (Local Binary Pattern) [35]. According to [34], a group is defined as two or more people who are spatially and temporally close to each other and exhibit a similar motion pattern. Based on this definition, [34] detects the groups in the scene. In group tracking, it evaluates group creation, update, split/merge of groups and group termination. Events are identified HMM based approaches with using knowledge extracted in the previous steps.

Similar to [34], [36] proposes a method for group tracking and behavior analysis in long video sequences in the underground railway station. Objects are extracted by using foreground detection and blob extraction. Then, it first tracks these extracted objects and group formation based on the definition on [34]. Once groups are identified, group tracking is performed with the method proposed in [34].

The most important part in group tracking is to evaluate interaction and grouping events (merge, split). Different approaches are used in the state-of-the art in order to perform this duty. [37] develops a group dynamical model and combines with interaction model based on Markov Random Fields (MRF). Tracking is performed by using Markov Chain Monte Carlo (MCMC) – Particle algorithm.

According to [38], a group is defined as the set of individuals who are spatially close to each other and have similar velocity and direction. In [38], Kalman filter is used as the tracking engine and it embeds rules of merge and split events into this framework. In order to handle a group with different velocities, it modifies the Probability Density Association (PDA) estimator so that it can handle this situation.

[39] proposes a group tracking method based on MHT in order to address the problem where maintaining the state of individuals is intractable. In [39], recursive MHT is selected to partition tracks to groups and associate observations to tracks. Split and merge events are handled by the multi hypothesis model.

There are also different approaches that exist in the state-of-the art for group tracking. [40] uses a genetic algorithm while [41] uses GM-PHD filter in tracking framework. Also, group tracking is used as a helper to crowd analysis methods. [42] proposes a crowd counting method based on group tracking. Crowd is divided into groups and it uses local features to count the number of people in group. Finally, it calculates crowd count as the sum of number of people in each group. During group tracking, it also handles the merge and split events.

### 2.2.3 Individual and Group Tracking

In the past few years, there has been a growing interest in handling individual and group tracking problems in a single framework [7, 28, 43].

[7, 43] propose a Decentralized Particle Filter (DPF) [44] based method and object state is decomposed into two sub-states: a group label, to which the individual belongs, and individual velocity and position. [7, 43] factorize the posterior distribution as follows:

$$p(Z_t, X_{0:t}|y_{0:t}) = p(Z_t|X_{0:t}, y_{0:t})p(X_{0:t}|y_{0:t}) \tag{2.4}$$

In this factorization, individual tracking result is also used in group state estimation shown in Figure 4.

[7, 43] redesign DPF [44] to track individuals and detect individual belonging group as shown in Algorithm 1.

**Algorithm 1: The DPF Algorithm in [7, 43]**

**INPUT:** Samples $\{X^{(i)}_{0:t}\}_{i=1,2..Nx}$ and $\{Z^{(i,j)}_{0:t}\}_{i=1,2..Nx,\ j=1,2...Nz}$. The $(i, j)$ means that each $i$ particle describing $X$ we have $N_z$ particles for describing $Z$.

1) Approximation of $p(X_{0:t}|y_t)$ through importance weights
   $$w_t^i \propto \frac{p_{Nz}(y_t|X^{(i)}_{0:t}, y_{0:t-1})p_{Nz}(X_t^{(i)}|X^{(i)}_{0:t-1}, y_{0:t-1})}{\boxed{\pi(X_t^{(i)}|X^{(i)}_{0:t-1}, y_{0:t-1})}}$$
2) Resample $\{X_t^{(i)}, Z_t^{(i,j)}\}$ according to $w_t^i$
3) Approximation of $p(Z_t|X_{0:t}, y_{0:t})$ through importance weights
   $$\overline{q}_t^{(i,j)} \propto \boxed{p(y_t|X_t^{(i)}, Z_t^{(i,j)})}$$
4) Generate $X_{t+1}^{(i)}$ according to $\pi(X_{t+1}|X^{(i)}_{0:t}, y_{0:t})$
5) Approximation of $p(Z_t|X_{0:t+1}, y_{0:t})$ through importance weights
   $$q_t^{(i,j)} \propto \overline{q}_t^{(i,j)}\boxed{p(X_{t+1}^{(i)}|X_t^{(i)}, Z_t^{(i,j)})}$$
6) Resample $Z_t^{(i,j)}$ according to $q_t^{(i,j)}$
7) Generation of particles $Z_{t+1}^{(i,j)}$ according to proposal $\boxed{\pi(Z_{t+1}|X^{(i)}_{0:t+1}, Z_t^{(i,j)})}$

**OUTPUT**: Importance sampling approximations of $X_{t+1}$ and $Z_{t+1}$.

Probability distributions shown in box are redesigned to track individuals and assign the group label for that individual as follows:

**Individual Proposal** $\pi(X_t^{(i)}|X^{(i)}_{0:t-1}, y_{0:t-1})$: This distribution models the individual motion model. To approximate this model, two sources of information are used as follows:

$$\pi(X_{t+1}|X_{0:t}, y_{0:t+1}) = \alpha\pi(X_{t+1}|X_t) + (1-\alpha)\pi_{det}(X_{t+1}|X_{0:t}, y_{0:t+1}) \quad (2.5)$$

In first part $\pi(X_{t+1}|X_t)$, locally linear dynamics with Gaussian noise is used as follows:

$$x_{t+1}^k = Ax_t^k + n \; with \; A = \begin{bmatrix} 1 & 0 & T & 0 \\ 0 & 1 & 0 & T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2.6)$$

where $n$ is the random Gaussian noise, $T$ is the time interval.

In the second part $\pi_{det}(X_{t+1}|X_{0:t}, y_{0:t+1})$, it uses a detector to estimate state vector of individuals. Parameter $\alpha$ is set once and kept fixed during the experiment.

**Joint Observation Distribution** $p(y_t|X_t^{(i)}, Z_t^{(i,j)})$**:** The aim in this distribution is to find the most similar template to the tracked object.

$$p(y_t|X_t, Z_t) \propto p(Z_t|X_t)p(y_t|X_t) \tag{2.7}$$

$p(\boldsymbol{X_t}|\boldsymbol{y_t})$ is the standard particle filter technique to find the most similar individual, $p(\boldsymbol{Z_t}|\boldsymbol{X_t})$ uses cluster validity method [45] which finds the closest group cluster to that individual.

$$p(y_t|X_t) = \exp(-\vartheta_d d(f(y_t, x_t), \tau) \tag{2.8}$$

where *d* is the distance between feature of $f(y_t, x_t)$ extracted from the current bounding box and $\tau$ template of individual. To find the group cluster for that individual, it uses following method

$$p(Z_t|X_t) \propto \exp(-\vartheta_{dc} dc(Z_t, X_t) \tag{2.9}$$

where *dc* is the Davies-Bouldin index [45] for cluster validity measurement.

**Joint Individual Distribution** $p(X_{t+1}^{(i)}|X_t^{(i)}, Z_t^{(i,j)})$: This distribution models the dynamics of individual taking into account group.

$$x_{t+1}^k = x_t^k + Bg_t^k + n \; with \; B = \begin{bmatrix} 0 & 0 & T & 0 \\ 0 & 0 & 0 & T \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, g_t^k = \frac{\sum_{l=1}^{K} x_t^l I(z_t^l == z_t^k)}{\sum_{l=1}^{K} I(z_t^l == z_t^k)} \tag{2.10}$$

where *I(.)* is the indicator function, $g_t^k$ is the position and velocity of the group the *k-th* individual belongs to.

**Joint Group Proposal** $\pi(Z_{t+1}|X_{0:t+1}^{(i)}, Z_t^{(i,j)})$: This distribution models the dynamics of the group.

$$\pi(Z_{t+1}|X_{0:t+1}, Z_t) = f\left(\prod\nolimits_g \pi\left(e_{t+1}^g \middle| X_{0:t+1}, g_t, \acute{g}_t\right), Z_t\right) \tag{2.11}$$

where $e^g \in \{Merge, Split, None\}$. To find which event *e* occurs, multinomial logistic regression is used as an offline learner. The learner is trained with the following features extracted from training data to find the correct event.

- The inter-group distance between the group $g$ and the closest group $\acute{g}$ considering the position and size ($d_{KL}$, Kullback-Leibler distance between Gaussians).
- The inter-group distance between the group g and the closest group $\acute{g}$ considering the velocities ($d_v$, Euclidean distance)

- Intra-group variance between positions of individuals in group $g$.

Different from [7], [43] propose a group formation and online inference mechanism based on the Dirichlet Process Mixture Models (DPMM). In [43], each individual is modeled as an observation from one of the infinitely many components of the Dirichlet Process mixture and this components is the group that individual belongs to. By this way, [43] propose an online inference mechanism which provides no need to explicitly model the group events such as the merge and split. This is one of the major advantages of [43]. However, there is no association and online learning mechanism for individual tracking. As a result of this, the number of the ID switches increases. Also, the individual tracking also affects the group state estimate. [7, 43] uses only the weighted color histogram in the group state estimate and other features like texture and shape are not considered. Also, the state model in [7, 43] provides the separation of individual state from group state, this increases the complexity of the system.

In addition to [7, 43], [28] proposes a tracking framework to analyze both the sparse and dense crowds by using the microscopic and macroscopic approaches. In the microscopic approach, an individual is evaluated separately and multi target tracking method is used as the tracking engine. In the macroscopic approach, which is used when there is a dense crowd; group tracking method is used instead of individual tracking method due to the occlusion problem. Figure 6 shows methodology proposed in [28].

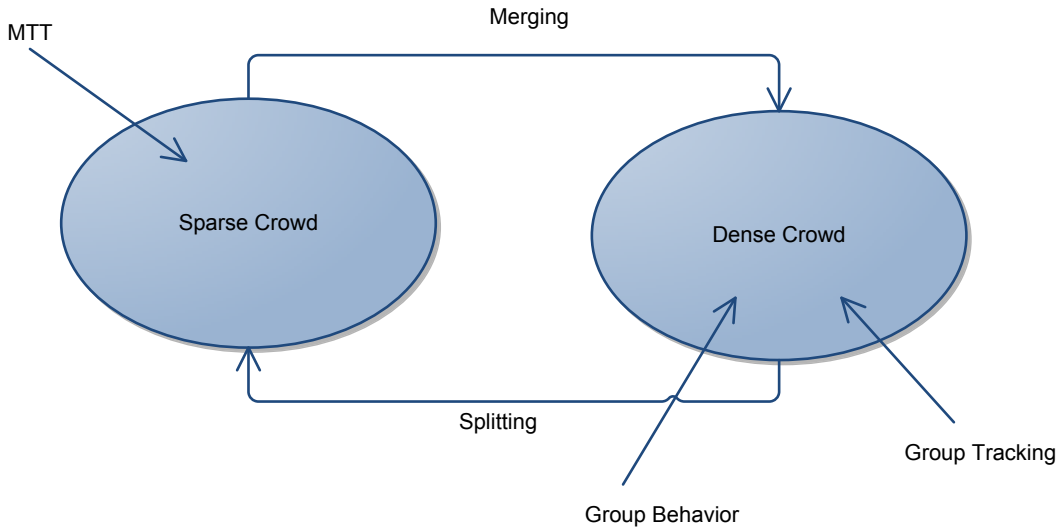



**Figure 6: The crowd analysis framework where both dense and sparse crowds can be analyzed in [28]**

In [28], people detection is performed by using foreground-background segmentation with Gaussian distribution. The Extended Kalman Filter is used as a tracking engine for detected people. Also, K-means clustering is used to analyze the detected clusters. The similarity between the clusters in K-means clusters is measured by the Euclidean distance and HMM is used to analyze the motion patterns of clusters.

This method uses the crowd density change to feed Hidden Markov Model (HMM) and estimate crowd activity. The major drawback in this method is that the test image consists of maximum of 8-10 people which are getting closer each other over time. However, dense crowd can be probably over ten people.

## 2.3 Related Works on Particle Advection

In this section, we provide the works about particle advection and its common use in the state-of-the art since we integrate the particle advection in our model to model the motion in the denser environments.

Particle advection is used to extract the motion flows of dense crowds and perform the stability analysis in crowds. [3] proposes a framework based on the particle advection in order to

extract the crowd flow and analyze the flow instabilities. In addition to [3], [4] performs the stability analysis from the motion flows for the behavior detection.

In addition to the flow analysis, the particle advection is used to detect abnormal events in the crowd. Individuals in the dense groups or crowds exhibits the similar motion patterns and as a result of this, the crowd has stable motion flow in normal cases. However, in the case of abnormal events, different type of the motion instability may be observed depending on the type of the abnormal event. [46] estimates the crowd velocity information with using the particle advection and uses this information in Gaussian Mixture Model (GMM) to detect abnormal events. In [47], the moving particles are treated as individuals and the interaction force is estimated by using SFM. Then, it obtains flows for every pixel in every frame and classifies each frame as normal or abnormal. [48] proposes a method to detect the dangerous events during mass events. This method extracts and analyzes the motion pattern of crowd by using the particle advection and detects the congestion in crowd. Similar to [46, 48], [49] performs the crowd segmentation and models the crowd motion by using the particle advection. Then, it analyzes the crowd motion to detect abnormal events. [50] proposes a method to optimize interaction forces obtained from the SFM. This method drifts the particle to the main motion areas and minimizes the interaction forces for abnormal event detection. [51] presents a unsupervised approach to cluster different behaviors and detect abnormal activities in the high density crowd by using Finite-Time Lyapunov Exponents (FTLE). Lyapunov Exponents is a measure of separation between infinitely close particles in infinite time. FTLE is calculated from the first position of particle and its last position by using particle advection. Then, it performs clustering operations by using an adaptive threshold method and uses these clusters to identify abnormal events.

Since the particle advection is used for the motion analysis, it can also be used for the person tracking. [52] treats the crowd as a set of particles and each particle represents individuals. In order to track specific individual, motion is modeled with a preference matrix which contains probabilities of moving to certain direction. By using this matrix, it performs tracking in high density crowds. [53] uses the particle advection to analyze the crowd motion pattern and uses it to track individuals in dense crowds. Although [52, 53] propose tracking approach for high density crowds, [54] proposes an individual tracking method in the low and high density crowds. In a low density crowd, the person detection is employed and then, data association

is performed based on Generalized Graph for each detection. In a high density crowd, individuals are tracked with the crowd flow modelling.

As result of literature review about particle advection, we observe that there is no explicit mechanism to extract the motion patterns for individual and group tracking approaches. In this thesis, we propose a motion extraction scheme by using particle advection explained in section 4.5.2.

# CHAPTER 3

# SOCIOLOGICAL BACKGROUND OF CONJOINT INDIVIDUAL AND GROUP TRACKING

In this section, we provide the sociological background and explain how they inspired us to develop the Conjoint Individual and Group Tracking.

In sociology, a group is defined in as two or more individuals who are related to each other by social relationships [55]. These social relationships provide people in the group share the similar characteristics and purposes. Major characteristics of the group are defined in [55] as *interaction*, *goals*, *interdependence*, *structure* and *unity* as shown in Table 1.

**Table 1: Characteristics of Groups**

| Characteristics | Explanation |
|---|---|
| **Interaction** | Task interactions among members |
| **Goals** | Groups have common purposes which facilitate the achievement of outcomes sought by the member |
| **Interdependence** | Each member in group influences and is influenced by each other member |
| **Structure** | Groups are organized as a pattern of relationships, roles, and norms. |
| **Unity** | Groups are cohesive social arrangements of individuals that perceivers consider as a whole. |

The group characteristics are not considered as separate. There are several relationships between these characteristics. One of these relationships is shown in Figure 7.

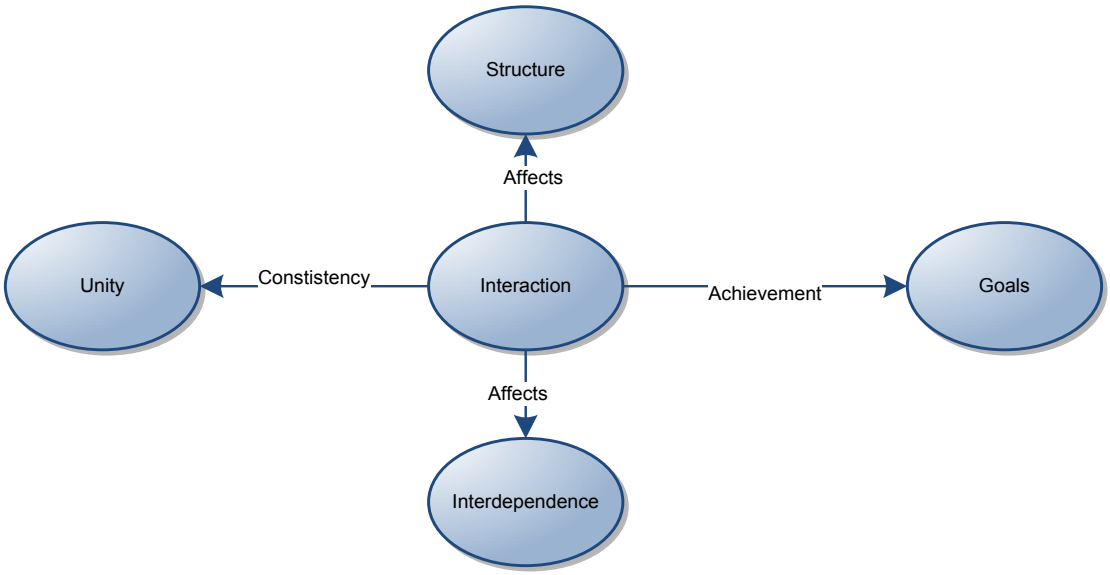



**Figure 7: Relationship between group characteristics**

In order to achieve the common goals, individuals in a group perform some actions and interactions but these interactions should be consistent and not break the unity feature of the group. Also, interactions affect the group formation, structure, and interdependency between the group members. As shown in Figure 7, since the interaction feature of the group is central in group analysis and it is measurable, most methods in the literature model the interaction in the group analysis and tracking. In [56], social force model is proposed to model the interactions between pedestrians in terms of repulsive and attractive forces called social forces. Repulsive forces are modelled according to the fact that individuals feel themselves uncomfortable when they get close to strangers and change their way and velocities when they face with obstacles and want to keep their distance to the border of obstacles. The social relationships between the individuals are used to model the attractive pulses. Based on these two forces, the interaction energy function is modelled for the multi target tracking.

Since the group is a dynamic entity and may shrink or grow as a result of split and merge events, not only the interaction between group members but also interaction between group members and other individuals should be taken into account for group tracking. Individuals in the group can be analyzed in terms of in-group and out-group measurements. In sociology, an in-group is defined as a social group to which a person identifies as being a member. By

contrast, an out-group is a social group with which an individual does not identify. [57] analyze the in-group and out groups effect on distributed teams and how they affect the effectiveness in partially distributed teams respectively. By inspiring from this definition, we define in-group as individuals who belong to that group and out-group is the individuals who are not part of that group. In order to track groups, the in-group measure and out-group measure are calculated for each individual.

In [58], the interaction is analyzed as a function of the distance and angular displacement between people. As a result of the unity feature, group members stay close to each other and perform similar angular directions in their movements. Also, in [59], the interaction possibility tends to increase when people get closer to each other. By using these features, the in-group and out-group measures in the multi-observation model of the CIGT tracking framework is calculated. In calculation of in-group and out-group weights, we inspired from the method in [58] where the social behaviors are modelled by means of individuals' velocities. According to [58], each pedestrian predicts the movement of other pedestrians and each pedestrian goes to a destination and try to avoid obstacles. By using this information, the energy function is built to model the individual's interaction for the multi target tracking as follows:

$$E_{ij}(v_i) = e^{-\frac{d_{ij}^2(v_i)}{2\sigma_d^2}} \tag{3.1}$$

where $E_{ij}$ is the interaction energy function between the person $s_i$ and person $s_j$, $d_{ij}$ is the distance between the person $s_i$ and person $s_j$ and $\sigma_d$ controls the distance to person to be avoided. The interaction is analyzed as a weighted sum of each person and weights are functions of current distances and angular displacements between people. For each person $s_i$ and person $s_r$, weight $w_r(i)$ is calculated as follows:

$$w_r(i) = w_r^d(i) w_r^{\varphi}(i) \tag{3.2}$$

$$w_r^d(i) = e^{-\frac{\|k_{ir}\|^2}{2\sigma_w^2}} \tag{3.3}$$

$$w_r^{\varphi}(i) = (\frac{1+\cos\varphi}{2})^{\beta} \tag{3.4}$$

where $w_r^d(i)$ and $w_r^{\varphi}(i)$ are the distance and angular displacement weights respectively, $k_{ir}$ is the distance between person $s_i$ and person $s_r$, $\sigma_w$ defines the radius of influence of other objects and $\beta$ controls the "peakiness" of weighted function. In CIGT framework, we calculate the observation measurements by means of angular displacement weights and distance similar to [58].

CIGT tracking framework does not only take account of interaction between individuals but also group structure. Group structure is affected by not only interaction. Number of people in group has a great effect on group structure. [55] states that the sparse and dense groups have different properties. As the group size increases, the group tends to become more complex and more formally structured. CIGT tracking model proposes a motion model for both sparse and dense crowds in terms of dynamic adjustment parameter and the particle advection method which can be used to analyze dense crowds [3, 4].

# CHAPTER 4

# CONJOINT INDIVIDUAL AND GROUP TRACKING WITH ONLINE LEARNING

In this thesis, we present a new individual and group tracking framework inspired from the sociological definitions and works. The proposed model uses a different factorization compared to the standard particle filter and evaluates not only visual features but also social interactions by means of in-group and out-group measurements.

The proposed framework is a complete framework in order to track both individuals and groups. It consists of four main parts as shown in Figure 8.

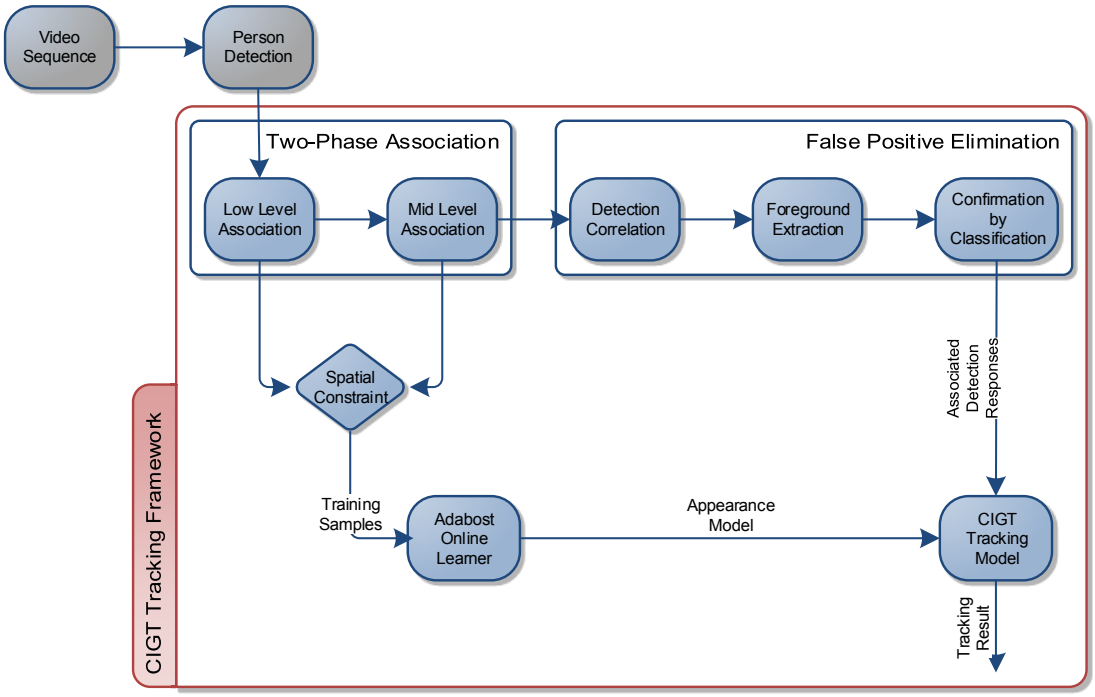



**Figure 8: CIGT Framework**

Two-phase association model contains low-level and mid-level associations and is used to associate individuals from previous frames. The results of association are fed to the online AdaBoost [60] with training samples obtained by applying spatial-temporal constraints. False positive elimination methods are used to reduce the false positive detections from detection responses. The core part of our framework is the CIGT Tracking Model which incorporates the multi-observation model, the motion model and the particle resampling phases.

## 4.1 Feature Selection

Feature selection is one of the most important steps in detection, association and tracking problems. Good features should be discriminative and efficient to compute in the visual tracking. In order to build strong appearance model, we need to select good features and features should be enough to identify object between each other.

An object can be described by means of its texture, shape and color features. In CIGT framework, we prefer to choose discriminative appearance model [5] since it provides a strong model to differentiate objects. Texture, shape and color features are described with region covariance [61], HOG [62] and color histogram.

In this section, we first provide information about appearance model representation and features used in CIGT framework. Then, we explain how to measure similarity between these features.

### 4.1.1 Appearance Model Representation

In CIGT framework, we use region covariance matrices for texture [61], RGB color histogram for color and HOG [62] feature for shape of tracked objects to build appearance model.

There are different methods in literature for texture analysis. According to [63], texture analysis methods can be categorized as statistical methods, structural methods, model-based methods and transform-based methods.

The Gray Level Cooccurance Matrix (GLMC) [64] is a statistical method for texture analysis and based on second-order statistics on gray scale image. The Run Length Matrix (RLM)

[65] uses higher-order statistics to define texture descriptor. Structural methods identify textures by means of composition of well-known shapes like rectangle, circle, line. [66] propose a structural method for texture analysis and uses different shapes of structuring elements. While these type of texture analysis methods can describe regular and static objects, they are not suitable for object tracking since the shape of tracked object may change in time. Model-based approaches are based on pixel values, for which Autoregressive (AR) [67] and Local Binary Pattern (LBP) [68] are some examples. The transform-based methods convert images from spatial domain to frequency domain. [69] uses Fast-Fourier Transform (FFT) while [70] uses Wavelet transform and Gabor filter for texture analysis.

In the proposed method, we utilize the region covariance [61] as a texture descriptor which proposes a statistical method for texture analysis. Unlike GLMC, the region covariance uses both first-order and second-order statistics to define texture. Also, it shows good performance in discriminative appearance model [5] for object tracking and the computational cost is very low.

Texture information can be described with a descriptor based on region covariance matrices of image features in [61]. The texture descriptor for the region *R* corresponding to the covariance matrix is defined as:

$$C_R = \frac{1}{n-1}\sum_{k=1}^{n}(z_k - \mu)(z_k - \mu)^T \tag{4.1}$$

$$z_k = \left[\frac{\partial I}{\partial x}\ \frac{\partial I}{\partial y}\ \frac{\partial^2 I}{\partial x^2}\ \frac{\partial^2 I}{\partial y^2}\ \frac{\partial^2 I}{\partial x \partial y}\right]^T \tag{4.2}$$

where $z_k$is the vector containing first and second derivatives of image at *k-th* pixel in the region *R*, $\mu$ is the mean vector over *R*, *I* is the grey-scale image patch, and *n* is the number of pixels.

Region covariance has several advantages. The noise in the region is mostly filtered out with an average filter during covariance computation. Also, it does not contain any information regarding orientation or number of point. Because of this feature, region covariance provides scale and rotation invariance over regions in different image.

Histogram is a widely used method to describe color representation of objects. In CIGT framework, we prefer to use RGB color space because of its simplicity and versatility and it shows good performance in appearance representation in [5]. Histogram is calculated with 8 bins for each channel. Then, these three vectors are concatenated to form single 24-element vector $f_{RGB_i}$ as a color representation of object.

HOG is a common feature to describe the shape of an object and used in human detection methods [71, 72, 73]. In CIGT framework, we employ HOG features [62] to extract shape information. A HOG feature $f_{HOG_i}$ is extracted over the region *R* with 8 orientations bins in 2×2 cells.

Finally, the appearance model for object $T_i$ can be written as:

$$A_i = \{ f_{RGB_i}, f_{HOG_i}, C_i \} \tag{4.3}$$

### 4.1.2 Similarity Measurements Between Features

Given texture, shape, and color descriptors above, we can calculate the similarity measures between regions for each feature.

We employ correlation coefficient for both color and HOG features due to its simplicity as follows:

$$p(f_x, f_y) = \frac{\sum_{i=1}^{n}(f_{x_i} - \overline{f_x})(f_{y_i} - \overline{f_y})}{\sqrt{\sum_{i=1}^{n}(f_{x_i} - \overline{f_x})^2 \sum_{i=1}^{n}(f_{y_i} - \overline{f_y})^2}} \tag{4.4}$$

where $p(f_x, f_y)$ is the correlation coefficient score between feature vectors $f_x$ and $f_y$, $\overline{f_x}$ and $\overline{f_y}$ are the mean value for feature vectors $f_x$ and $f_y$, $f_{x_i}$ and $f_{y_i}$ denote *i-th* element of feature vectors $f_x$ and $f_y$.

The similarity measurement for covariance matrices is more complex and described in [61] as follows:

$$\sigma(C_i, C_j) = \sqrt{\sum_{k=1}^{5} ln^2(\lambda_k(C_i, C_j))} \tag{4.5}$$

where $\lambda_k(C_i, C_j)$ are the generalized eigenvalues of $C_i$ and $C_j$ and computed from

$$\lambda_k C_i x_k - C_j x_k = 0 \quad k = 1, \dots, 5 \tag{4.6}$$

and $x_k \neq 0$ are the generalized eigenvectors.

## 4.2 Two-Phase Association

Two-phase association is performed on detection responses and aims to identify individuals from previous frames. The first phase aims to identify individuals on consecutive frames while the second phase provides a long-term association.

Two-phase association also helps to feed AdaBoost Online Learner with training samples which are obtained by applying spatial constraint on association result.

### 4.2.1 Low-Level Association

In this phase, we use dual-threshold method in [11]. The affinity score of each detection response is calculated by multiplying of measurements of position, size and color histogram for each object [5]. Then, affinity score matrix *S* is formed. Each element of this matrix defines the affinity score between objects in the previous frame and detection responses in the current frame. In this phase, we consider detection responses $r_i$ and $r_j$ belong to the same tracked object if the following conditions are satisfied:

$$\begin{aligned} \{x, y |\, S(i,j) > \theta_1 and\ |S(i,j) - S(x,j)| > \theta_2 \\ and\ |S(i,j) - S(i,y)| > \theta_2, x \neq i, y \neq j\} \end{aligned} \tag{4.7}$$

where $S(i,j)$ is the affinity score between detection responses $r_i$ and $r_j$, $\theta_1$ and $\theta_2$ are two thresholds used in [13]. In our CIGT framework, these thresholds are tuned according to features used in affinity score calculation. Also, $\theta_2$ is tuned according to $\theta_1$.

Dual-threshold strategy provides conservative and biased way to link only reliable associations. The low-level association does not resolve ambiguity of conflicting pairs and only performed in consecutive frames; therefore, mid-level association is used to solve these problems.

### 4.2.2 Mid-Level Association

Mid-level association is considered to be a long-term association method and performed only for only detections that cannot be associated in the low-level association phase. Unlike low-level association, we consider not only appearance model but also spatial information that belongs to objects. Therefore, the appearance similarity score and motion score are calculated separately for detection responses.

In appearance similarity score, we use a discriminative appearance model proposed in [5]. Object's appearance representation explained in 4.1 is modeled by means of texture, shape and color. Therefore, in CIGT framework, we calculate appearance similarity score as follows:

$$S_a = G\big(\sigma(C_i, C_j)\big) G\left(-1 + p(\, f_{RGB_i}, \, f_{RGB_j})\right) G\left(-1 + p(\, f_{HOG_i}, \, f_{HOG_j})\right) \tag{4.8}$$

where $G(.)$ is the zero mean Gaussian function, $p(.)$ is cross correlation coefficient function, and, $\sigma$ is distance function between two region covariance matrix.

Once appearance score is obtained, spatial information belonging to the object is used in mid-level association by computing motion similarity measure. Inspired from [17], we calculate the motion similarity score as function of distance between the position estimated with linear motion model $p^e_{t+\Delta t}$ and the real position $p^a_{t+\Delta t}$. Figure 9 shows calculation of motion similarity score.

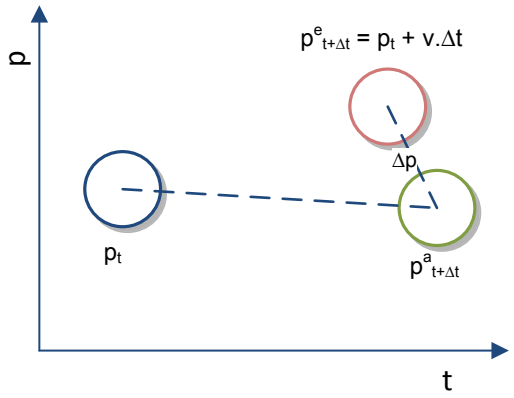



**Figure 9: Linear motion model to estimate position. The horizontal axis denotes the time, vertical axis denotes the position**

In CIGT, we keep velocity of detection associated to tracked object. Therefore, we can calculate the estimated position by using linear motion model and position difference as follows:

$$p^e_{t+\Delta t} = v_t * \Delta t + p_t \tag{4.9}$$

$$\Delta p = p^a_{t+\Delta t} - v_t * \Delta t - p_t \tag{4.10}$$

where $v_t$ is the individual velocity, $p_t$ is the individual position at frame $t$, $\Delta t$ is the frame difference, $p^e_{t+\Delta t}$ is the estimated position, and $p^a_{t+\Delta t}$ is the actual position at frame $t + \Delta t$. Then, we compute the motion similarity score as follows:

$$S_m = G(\Delta p, \sum_p) \tag{4.11}$$

where $G(.,\sum)$ is the zero mean Gaussian function. Finally, the probability of association is calculated with appearance and motion similarity scores as follows:

$$P_s = S_a \times S_m \tag{4.12}$$

## 4.3 False Positive Elimination

CIGT proposes a hierarchical method to reduce the number of false positives in the detection results. As shown in Figure 8, CIGT performs the elimination of false positive in three phases: detection correlation, foreground extraction [74] and confirmation by classification [9].

The detection correlation is proposed in CIGT framework and basically eliminates the false positive detections after two-phase association. The key point in detection correlation is to correlate detections not associated by low-level and mid-level associations to another uncorrelated detection or detections associated by low-level or mid-level associations. Most of the detections are identified by two-phase association and our aim is to eliminate the detection of suddenly appearing objects, not satisfying spatial constraint or intersecting with others. Therefore, we compute minimum intersection ratio between the two detections in order to select redundant detections as follows:

$$I_s = \min\left(\frac{A_i}{A_1}, \frac{A_i}{A_2}\right) \tag{4.13}$$

where $A_i$ is the intersection area of the detection areas $A_1$ and $A_2$. The other feature of dual designated detection is the similarity of other detection that is already associated by two-phase association. Therefore, we compute the appearance similarity measure between two detections by means of histogram, HOG features and size as follows:

$$A_s = \sqrt[3]{P_{hist} P_{HOG} P_{size}} \tag{4.14}$$

where $P_{hist}$, and $P_{HOG}$ are histogram, HOG cross correlation distance and $P_{size}$ is the size distance between two detections. Finally, the two-threshold method is applied as follows:

$$DC_s = \begin{cases} 0\,, if\ A_s > \delta_1 \text{ and } I_s > \delta_2 \\ 1, \text{otherwise} \end{cases} \tag{4.15}$$

where $DC_s \in \{0,1\}$ is the detection correlation result. 0 means not correlated and 1 means correlated. $\delta_1$ and $\delta_2$ are the appearance similarity and intersection thresholds. The appearance similarity threshold is tuned according to features used in appearance model. Intersection threshold is set to 0.8 since correlated detections are close enough to each other. Correlated detections are labeled as a false positive and eliminated.

In the second step, we use foreground extraction to eliminate false positive detections. In CIGT framework, we choose multi-layer background subtraction proposed in [74] which uses local texture feature represented by local binary patterns and photometric invariant color measurement in RGB color space. Since these texture and selected color feature are illumination invariant, [74] can handle local illumination changes. Therefore, it provides less noisy results as shown in Figure 10.

**Figure 10: Multi-Layer Background Subtraction [74] Results**

If detection is not identified by two-phase association and foreground area in detection is small, detection is labelled as false positive.

In the last stage, confirmation-by-classification method in [9] is used. According to [9], if tracked object is not updated by detection responses for a certain number of frames, then tracked object is eliminated. Detection correlation and foreground extraction are designed to eliminate false positive from current detection responses. However, Confirmation-by-Classification [9] method aims to eliminate false positives that are labelled as a true positive in previous frames.

## 4.4 AdaBoost Online Learning Model

In CIGT framework, we use the Real-AdaBoost algorithm [60] as a learning model to determine appearance similarity between two instances and compute confidence score of similarity measure. Our model takes two instances as the input and performs a binary classification for these instances. Confidence score is calculated by Equation 4.8.

AdaBoost algorithm creates one strong hypothesis from several weak hypotheses. In CIGT framework, the training samples are collected based on method in [5] by using spatial constraint based on the fact that object cannot be at two locations in the same frame. After two-phase association, we collect training samples from only associated objects. Negative samples are created from two associated objects which are spatially separated from each other. For instance, as shown in Figure 11, four tracklets are associated by two-phase association. $T_2$ and $T_3$ are too far away from each other and these detections corresponding to $T_2$ and $T_3$ are used to create negative samples. Also, we build discriminative set for each tracklet by using method in [5]. For tracklet $T_i$ , $D_j$ is built from $T_j$ where $T_i$ is far away from $T_j$ in spatial domain. By using spatial constraint and discriminative set, we collect negative and positive samples and build weak hypothesis.

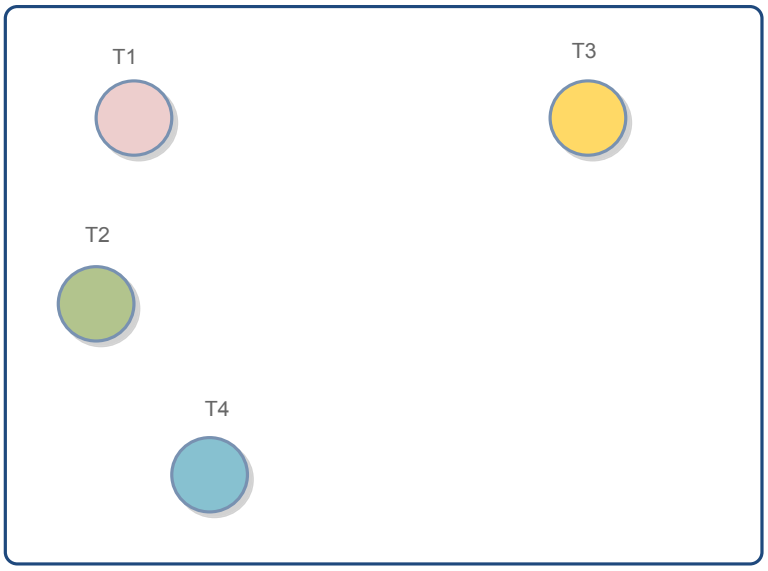



**Figure 11: Spatial-Temporal Constraint and collecting training samples**

Weak hypothesis consists of region covariance, HOG features and histogram similarity measures as follows:

$$h(A_i, A_j) = \left[p\left(f_{RGB_i}, f_{RGB_j}\right)\, \sigma(C_i, C_j)\, p\left(f_{RGB_i}, f_{RGB_j}\right)\right] \tag{4.16}$$

where $h(A_i, A_j)$ is the weak hypothesis, $p\left(f_{RGB_i}, f_{RGB_j}\right)$, $\sigma(C_i, C_j)$, and $p\left(f_{RGB_i}, f_{RGB_j}\right)$ are the histogram, region covariance, and hog similarity measures between appearance mod-

els $A_i$ and $A_j$. By using these weak hypotheses, AdaBoost algorithm builds a strong classifier as follows:

$$H(A_i, A_j) = \sum_{k=1}^{K} \alpha_k h(A_i, A_j) \tag{4.17}$$

where $\alpha_k$ are the parameters to be estimated, $H$ is a strong hypothesis, and $h$ is a weak hypothesis. In our model, we aim to estimate $\alpha_k$ parameters by minimizing the loss function. The loss function in AdaBoost is defined as follows:

$$Z = \sum_i w_i^0 e^{-y_i H(x_i)} \tag{4.18}$$

where $Z$ is the loss function, $w^0$ is the initial weight. Our aim is to find a strong hypothesis $H(x)$ that minimizes loss function $Z$, where $H(x)$ is obtained sequentially adding new weak hypothesis. Our goal is to minimize $Z$ at *k*-th round as follows:

$$Z_k = \sum_i w_i^k e^{-y_i \alpha_k h_k(x_i)} \tag{4.19}$$

At each round, we update the sample weights with respect to $\alpha_k$ and $h_k$. Algorithm 2 summarizes the learning appearance model.

**Algorithm 2: AdaBoost Online Learning Appearance Model Algorithm**

***Input*:** $\boldsymbol{\beta^+ = \{(x_i, +1)\}}$**:** Positive samples
$\boldsymbol{\beta^- = \{(x_i, -1)\}}$**:** Negative samples
$\boldsymbol{F = \{h(x_i)\}}$**:** Weak hypothesis pool

1. **Set** $\boldsymbol{w_i = \frac{1}{|\beta^+ \cup \beta^-|}}$
2. **for** *t* = 1 to *T* **do**
3. **for** *k* = 1 to *K* **do**
4. $\boldsymbol{r = \sum_i w_i y_i h_k(x_i)}$
5. $\boldsymbol{\alpha_k = \frac{1}{2} In(\frac{1+r}{1-r})}$
6. **end for**
7. Choose $\boldsymbol{k^* = arg\ min_k \sum_i w_i e^{-\alpha_k y_i h_k(x_i)}}$
8. Set $\boldsymbol{\alpha_k = \alpha_{k^*}}$ and $\boldsymbol{h_t = \alpha_{k^*}}$
9. Update $\boldsymbol{w_i = w_i e^{-\alpha_k y_i h_k(x_i)}}$
10. Normalize $\boldsymbol{w_i}$
11. **end for**

***Output***: $\boldsymbol{R_a = H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))}$

Once classification is performed and AdaBoost binary classification result is obtained, confidence score is calculated as follows:

$$S_c(A_i, A_j) = \begin{cases} -1, & if\ R_a = -1 \\ e^{-cc\left(f_{RGB_i}, f_{RGB_j}\right)\sigma(C_i, C_j)cc\left(f_{RGB_i}, f_{RGB_j}\right)}, & if\ R_a = +1 \end{cases} \quad (4.20)$$

where $S_c(A_i, A_j)$ is confidence score, $R_a$ is the AdaBoost binary classification result.

## 4.5 Conjoint Individual and Group Tracking Model

Group has a structured and dynamic entity that may grow or shrink with merge or split events. Our starting point is that a person is actually a one-person group. Consequently, if we model the particle filter for a group, then we can use this method for the individual. However, standard particle filter model does not provide such a dynamic properties. In this thesis, we propose new tracking mechanism based on particle filter so that we can track both groups and individuals. Figure 12 shows CIGT architecture.

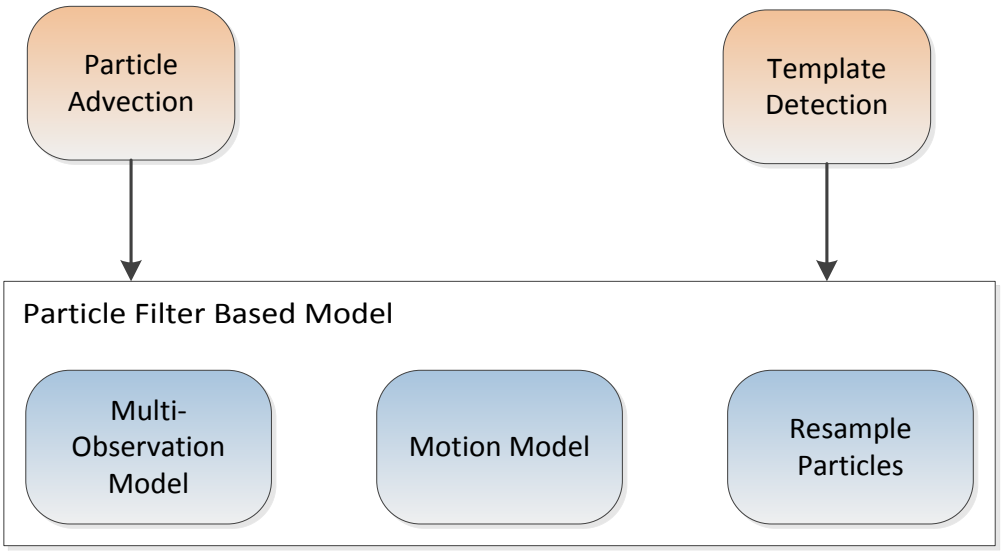



**Figure 12: CIGT Architecture**

In our proposed framework, we use Particle advection [3, 4], and Template Detector to extract information and input into Particle Filter based Model.

In CIGT, we propose a new observation model named multi observation shown in Figure 13. This model provides us to detect merge and split events for tracked group.

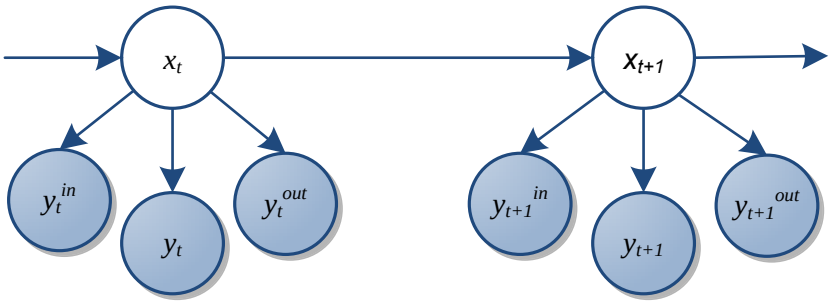



**Figure 13: CIGT Multi-Observation Model**

As described in section 3, in-group is defined as a group that a person identifies him/herself as being a member and out-group is defined as a group that individual does not identify him/herself as a member in sociology. In CIGT, we define in-group as individuals who belong to that group and out-group is the individuals who are not part of that group and apply this definition into a multi observation model as shown in Figure 13, where the observation model consists of in-group ($y_t^{in}$), out-group ($y_t^{out}$) observations and similarity measure ($y_t$). In CIGT, each person area in group has in-group measure and out-group measure. In-group measure is used to identify whether person is part of group or not and out-group measure identifies whether person is part of group or not. This model provides us dynamic model by detecting merge and split events for the tracked group and described in the following section.

### 4.5.1 Multi-Observation Model

Since a group may grow or shrink with merge and split events, we need to model these group dynamics in order to track groups. People in a group share similar characteristics as a result of unity property. Also, [59] states that interaction possibility increases while people get closes to each other. Since group member has interaction between each other and shares the similar properties, they are expected to have similar velocities and keep close to each other.

The main idea behind the multi-observation model is to decompose standard particle filter observation model into disjoint observations. Since individual similarity and interaction with other people are disjoint, observation can be decomposed into as the person similarity and interaction measure with the other people as follows:

$$P(y_t|x_t) = p(y_t^i, y_t^s|x_t) = p(y_t^i|y_t^s, x_t)p(y_t^s|x_t) \tag{4.21}$$

where $y_t^i$ is the observation for interaction with other people, $y_t^s$ is the individual similarity observation. Due to the independency of similarity and interaction observations between each other, we obtain:

$$P(y_t|x_t) = p(y_t^i|x_t)p(y_t^s|x_t) \tag{4.22}$$

The multi-observation model also decomposes interaction observation into in-group and out-group observations. In-group observation evaluates the interaction between each person in specific group and other people in this group while out-group observation performs the interaction evaluation between each person in specific group and other people not belonging to this group. Since the evaluation for each person in group is performed with two disjoint set of people, we can decompose interaction observation as follows:

$$p(y_t^i|x_t) = p(y_t^{in}, y_t^{out}|x_t) = p(y_t^{in}|y_t^{out}, x_t)p(y_t^{out}, x_t) \tag{4.23}$$

$$p(y_t^i|x_t) = p(y_t^{in}|x_t)p(y_t^{out}, x_t) \tag{4.24}$$

After performing these decompositions, observation model becomes:

$$P(y_t|x_t) = p(y_t^{in}|x_t)p(y_t^{out}, x_t)p(y_t^s|x_t) \tag{4.25}$$

In-group observation $p\left(y_t^{in}, x_t\right)$ is a measure of degree of belonging to a specific group. Out-group observation $p(y_t^{out}, x_t)$ is a measure of degree of not belonging to a specific group. Since group member share similar characteristics and interact each other due to unity and interaction features, it is expected from group member to exhibit similar velocity, motion direction and close to each other.

In multi-observation model, we use the direction similarity of individuals, and closeness degree in order to compute in-group and out group weights. Direction similarity can be meas-

ure as a function of angle between two motion vectors belonging to two individuals. Therefore, direction similarity and closeness degree are computed as follows:

$$w_{x,y}^{\theta} = \frac{1 + \cos\theta}{2} \tag{4.26}$$

$$w_{x,y}^{d} = e^{-\frac{d}{\min_{s}(s.width, s.height)}} \tag{4.27}$$

where $\theta$ is the angle between individual x and y, $d$ is the distance between individual x and y, and $s$ is the list of all individuals detected in first frame and minimum value of width and height values belonging to $s$ is used to normalize distance between individuals. Interaction weight is computed as follows:

$$w_{x,y}^{i} = w_{x,y}^{\theta} w_{x,y}^{d} \tag{4.28}$$

Initially, in-group and out-group measures are set to $w_i$ such that $w_i < 0$ and updated as follows:

$$w_{x,y}^{g} = \begin{cases} 0, & if\ w_{x,y}^{d} > w^{d}\ \ or\ w_{x,y}^{\theta} > w^{\theta} \\ w_{x,y}^{i}, & otherwise \end{cases} \tag{4.29}$$

where $x \in P^{in}$, g $\in \{in, out\}$, $y \in P^{in}$ and $x \neq y$ if $g = in$ or $y \in P^{out}$ if $g = out$ , $P^{in}$ is the set of people in the group, $P^{out}$ is the set of people in out-group, $w_{x,y}^{g}$ is the weight for particles belonging to individual x with respect to individual y, $d_{x,y}$ is the Euclidian distance between individual x and y for all particle, $w^{d}$ and $w^{\theta}$ are the threshold values of closeness degree and direction similarity. Since $w_{x,y}^{d}$ is computed with respect to the normalized distance, we set $w_{x,y}^{d} = e^{-1}$. Since for direction similarity, people in group should not head to opposite directions, implying that the angle between individual's directions vectors $\theta$ should not be greater than $\pi/2$. Therefore, the $w^{\theta}$ is computed by taking $\theta = \pi/2$ so $w^{\theta} = 0.5$.

After the calculation of in-group and out-group weights, it is needed to refine the in-group and out-group weights for each particle since some individuals connect to others indirectly. Therefore, we update the weights if $w_{x,y}^{g} = 0$ but there is an individual z for which $w_{x,z}^{g} > w^{\theta}w^{d}$ and $w_{z,y}^{g} > w^{\theta}w^{d}$. In this case, we set $w_{x,y}^{g} = w_{z,y}^{g}$ as x is connected to y through z..

For example, in Figure 14, for persons *A*, *B* and *C* where $w^{g}_{A,B} > 0$, $w^{g}_{A,C} = 0$ and $w^{g}_{B,C} > 0$. In this case, we can say that *A* is connected to *C* through *B* and update the weight between *A* and *C* accordingly

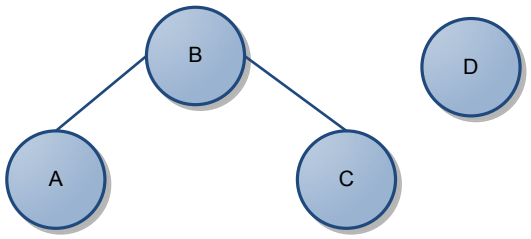



**Figure 14: Undirected Graph for Weight Refinement in CIGT**

After creating undirected graph for objects in group, we update weight table as follows:

| | **A** | **B** | **C** | **D** | | | **A** | **B** | **C** | **D** |
|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 1 | 1 | 0 | 0 | | **A** | 1 | 1 | **1** | 0 |
| **B** | 1 | 1 | 1 | 0 | → | **B** | 1 | 1 | 1 | 0 |
| **C** | 0 | 1 | 1 | 0 | | **C** | **1** | 1 | 1 | 0 |
| **D** | 0 | 0 | 0 | 1 | | **D** | 0 | 0 | 0 | 1 |

**Figure 15: Weight Table for Refinement process in CIGT**

Refinement is an iterative process and continues until there are no further connections between the nodes. As a result of this process, all edges between nodes are created and weights are calculated according to the weight table.

After refinement of all individual's in-group and out-group weights, event $e \in \{Merge, Split, None\}$ is decided as follows:

$$E(x; y, z) = \begin{cases} Split, if\ \ w^{in}_{x,y} = 0 \\ Merge, if\ w^{out}_{x,z} > 0 \\ None,\ \ otherwise \end{cases} \quad (4.30)$$

where $x, y \in P^{in}$, $z \in P^{out}$, $P^{in}$ is the set of people in the group, $P^{out}$ is the set of people in out-group. Function E detects the split event between individual x and y, merge event between individual x and z. The individual similarity observation $p(y_t^s|x_t)$ is calculated with L2-norm.

### 4.5.2 Motion Model

Similar to [7, 43], there are two sources used in the motion model of CIGT: particle advection and template detector. Basically, the static particles are created and distributed on the scene and then dense optical flow is calculated for each particle in order to analyze major motion in the scene. It is mainly used in flow extraction and stability analysis [3, 4] and has been shown to be an effective method to obtain motion information in crowded scenes. As template detector, correlation coefficient is used. Consequently, motion model in CIGT is formulated as follows:

$$\pi(X_{t+1}|X_{0:t}) = \alpha\pi_{Det}(X_{t+1}|X_t) + (1-\alpha)\pi_{PA}(X_{t+1}|X_{0:t}, y_{0:t+1}) \tag{4.31}$$

where $\pi_{PA}(X_{t+1}|X_t)$ and $\pi_{Det}(X_{t+1}|X_t)$ are hypothesizes that generate motion vector with the particle advection, and the template detector, respectively. The $\alpha$ is a parameter that determines the relative weights of the particle advection based model and the detector model. In CIGT, the linear dynamic model in [7, 43] is used.

$$x_{t+1}^k = Ax_t^k + n_g\,, \qquad A = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.32}$$

where A is the particle dynamics, $x_t^k$ is the group state at time t and $n_g$ is Gaussian noise. In CIGT tracking model, object state is composed of position and velocity for both groups and individuals, $x_t = [p_x\ p_y\ v_x\ v_y]$. The velocities $[v_x\ v_y]$ are calculated by using both particle advection and template matching.

One of the biggest challenges in particle advection model is to calculate object velocity. Since particles are distributed statically, we need to find reliable particles in order to compute velocity. Three successive steps are used to choose these reliable particles:

(i) Distance filter

(ii) Cross correlation with the optical flow

(iii) Forward-backward optical flow [19].

(iv) Unreliable particle elimination

Distance filter is used to choose the particles in or very close to the tracked object. In cross correlation with optical flow, these particles are moved by using forward Lucas-Kanade optical flow [75]. $\omega \times \omega$ rectangles are taken from the current and following frames. Then, template match measure is computed for each particle. We only keep particles whose matching score is high. Forward-backward optical flow method in [76] is applied for retained particles shown in Figure 16.

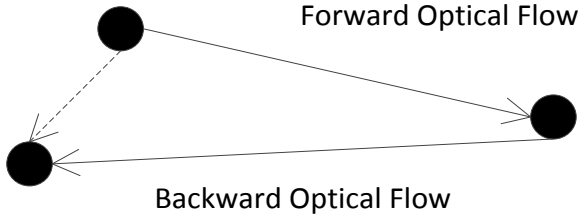



**Figure 16: Forward- Backward Optical Flow Validation**

In forward-backward optical flow method, particles are moved with forward optical flow, and then move back with backward optical flow. The distance between original position and forward-backward position of particle is computed as a distance error. We keep only particles whose distance error is very low. Finally, we propose unreliable particle elimination method in order to choose particle whose motion history is consistent. This method provides us to select correct particle when occlusion event occurs. In this method, velocity vector belonging to each racked object is kept for some frame and estimate similarity measure current velocity vector according to velocity vector history. We choose only particles whose velocity vector is consistent with tracked object's motion history. After choosing particles according these four methods, $\pi_{PA}(X_{t+1}|X_t)$ computes the velocity vector by using these particles. The template detector model $\pi_{Det}(X_{t+1}|X_t)$ calculates $[v_x\ v_y]$ by estimating the next state of the object with detection.

The detector model $\pi_{Det}(X_{t+1}|X_t)$ uses the correlation coefficient template matching and calculates $[v_x\ v_y]$. Template matching is defined as follows:

$$R(x,y) = \sum_{x',y'} \left( T(x',y') - \frac{1}{w.h} \sum_{x'',y''} T(x'',y'') \right) .I(x+x', y+y') \tag{4.33}$$

where *I* is the input image, *T* is the template to be found, *R* is the result of correlation coefficient template matching, *w* and *h* are width and height of the template respectively.

One of our contributions is the α parameter. Unlike [7, 43], α parameter is dynamic and changing with respect to the group density. When α parameter increases, contribution of detector to motion model increases as well. Therefore, $\alpha \sim m_{measure}$ and $m_{measure}$ is the matching measure. Also, $m_{measure}$ should be in [min, max] range. Figure 17 shows relationship between α and $m_{measure}$.

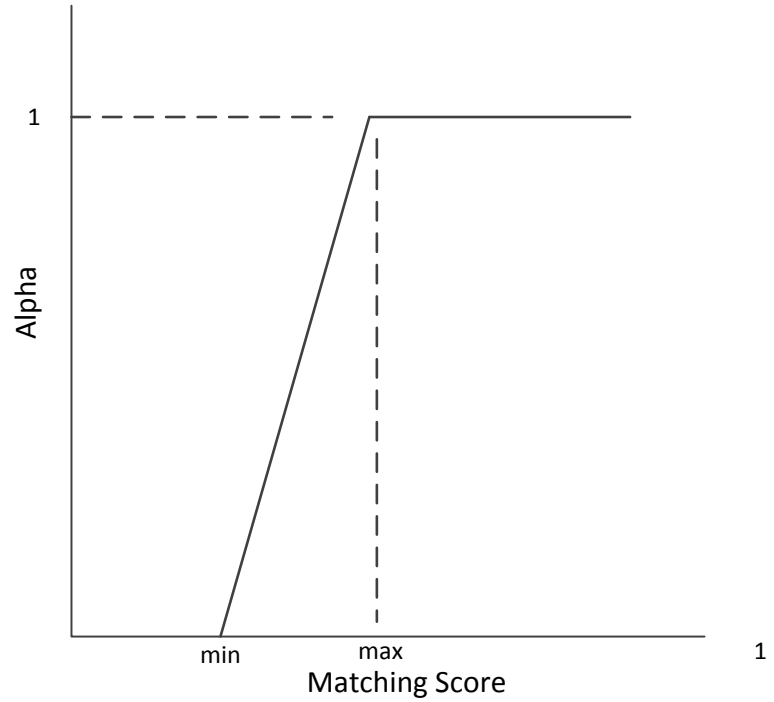
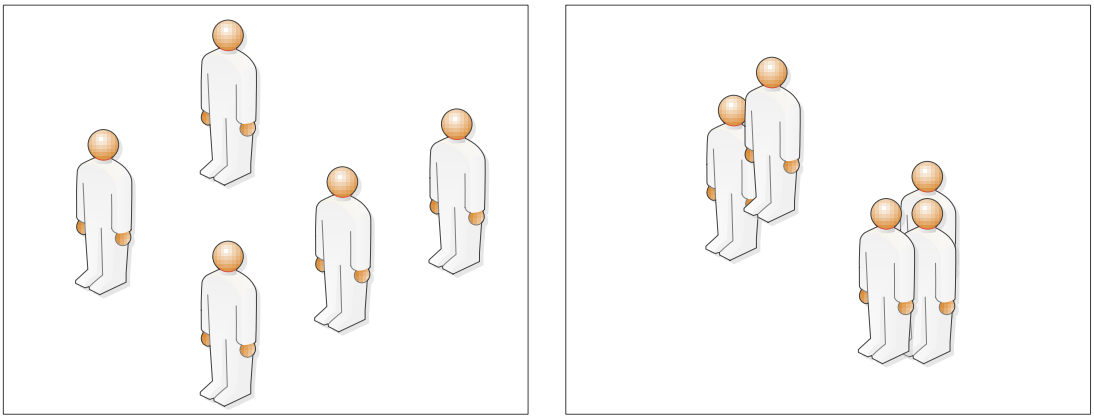



**Figure 17: Graph of Alpha versus Matching score**

If we formalize this graph, then we can obtain that

$$\alpha \sim \frac{m_{measure} - m_{min}}{m_{max} - m_{min}} \tag{4.34}$$

We know that particle advection will give us better results while number of people in group increases. However, we also need to take closeness between people in the group into account. If the number of people is high but people are not close and can be extracted separately, then detector gives us good results as well. Also, tracking individually will be more effective. Figure 18 shows these people state in group.



(a) Separated People (b) Occluded People

**Figure 18: People closeness types. (a) Separated People (b) Occluded People**

In order to address this case, we define complexity factor for group as follows:

$$c_g = \frac{n_{people}}{n_{blob}} \tag{4.35}$$

where $c_g$ is the complexity of the group, $n_{people}$ is the number of people in the group and $n_{blob}$ is the number of blobs in the group.

As mentioned earlier, particle advection will give us better result with increasing group complexity. Therefore, we can define $\alpha \sim {}^{1}/_{c_g}$. Our α parameter changes as follows:

$$\alpha = \frac{m_{measure} - m_{min}}{c_g(m_{max} - m_{min})} \tag{4.36}$$

where $m_{min}$ is the minimum acceptable matching measure and $m_{max}$ is the maximum matching measure, $m$ is the calculated matching measure between object and the detected template. Using $m_{min}$ and $m_{max}$ allows normalizing $\alpha$ parameter into the range [0-1].

### 4.5.3 Particle Resampling

The other important step other than motion model and observation model is the resampling of particles. In CIGT, we use a dynamic particle generation method that we developed.

In the standard particle filter, the number of particles is set once and does not change during tracking. However, since groups are dynamic entities and group sizes may change with merge and split events, particle size will not be sufficient while group size increases. CIGT tracking model proposes a dynamic particle generation method which changes the number of particles with respect to the number of detected persons.

Standard particle filter uses sequential importance, evaluates the tracked area as a single-piece and resamples the particles according to the best observation. Instead, we find the best observations for all person regions in the tracked group and resample the particles according to them. The number of particles is computed dynamically during tracking as follows:

$$p_{total} = p.n \tag{4.37}$$

where $p$ is the number of particles for one person, $n$ is the number of people in the tracked area and $p_{total}$is the number of particles to be used in CIGT.

### 4.5.4 State Estimate

State estimate in CIGT is performed for both groups and individuals. Confidence score in AdaBoost online learning model is used to compute similarity score for each individual in a group. Group similarity measure is computed as follows:

$$S_{est} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} AdaboostScore(p_i, q_j)}{N} \tag{4.38}$$

where $S_{est}$ is the state estimate for group or individual, and $N$ is the number of individuals in the group. Also, confidence score for all individuals in a group should be positive. The power of state estimation comes from online learning model and using not only color information but also texture and shape information in order to distinguish individuals. State estimate procedure is summarized in Algorithm 3.

**Algorithm 3: CIGT Framework State Estimation Algorithm**

**Input:** $\boldsymbol{G_1}$**:** Tracked Object in current frame**,** $\boldsymbol{G_2}$**:** Tracked Object in previous frames

```
if G1.size ≠ G2.size
  return 0
else
  N ← G1.size
end
Sest ← 0
For each individual p ∈ G1
 For each individual q ∈ G2
   S ← AdaboostScore(p, q)
   if S > 0
     Sest ← Sest + S
   else
     return 0
   end
 end
end
Sest ← Sest/N
```

**Output:** State estimation measure $\boldsymbol{S_{est}}$

# CHAPTER 5

# EXPERIMENTS AND RESULTS

In this chapter, we experimentally evaluate the proposed CIGT framework and compare the results against those of the state-of-art on different datasets. Recall that CIGT framework proposes a Multi-Observation Model which allows evaluating social interactions in groups with merge and split events. Also, CIGT framework evaluates the group crossing in event evaluation. We assume that the videos are captured by stationary cameras and datasets are chosen accordingly.

Our experiments consist of four main parts. In the first part, we conduct experiments to obtain results with individual and group tracking metrics and evaluate these results with other researches in state-of-art. Secondly, we investigate the effects of person detection errors on proposed framework by simulating the person detection with different error ratio. Thirdly, we evaluate the dynamic parameter used in motion model by comparing edge values of this adaptive parameter. In the final part, we provide the performance result of the proposed framework, discuss these results and make a suggestion about how to improve performance with respect to the memory utilization and time.

## 5.1 Dataset

One of the biggest challenges is the limited number of datasets which can be used to evaluate individual and group tracking. We selected the datasets to be used by considering following criteria:

(i) Dataset shall include scenarios with merge and split events.
(ii) Dataset shall include scenarios with dense groups.
(iii) Dataset can be used in comparisons with the state-of-art frameworks in the literature.

Under these considerations, we use the datasets shown in Table 2 in our experiments.

**Table 2: Datasets used in experiments**

| Dataset | Number of Videos |
|---|---|
| **FM Dataset Synthetic [7, 43]** | 25 |
| **FM Dataset Real [7, 43]** | 13 |
| **BIWI Dataset [58]** | 2 |
| **PETS 2009 [77]** | 6 |
| **Total** | **46** |

The proposed method is tested with 4 datasets: Friends Meet (FM) Synthetic and Real [7, 43], BIWI [58], and PETS 2009 [77]. FM and BIWI datasets include ground-truth information. In our experiments, we simulate the person detector for FM and BIWI datasets in same way with [7, 43] by generating detections from the ground-truth with a false positive and false negative of 20% and adding spatial Gaussian noise. Also, we used real a person detector [78] in order to evaluate PETS 2009 so that we can observe results with implemented person detector. For group evaluation in PETS 2009 dataset, we annotate group ground truth from individual group truth [77] and evaluation was performed with both these group and individual ground truth data.

The FM dataset [7, 43] consists of 53 sequences with 16286 frames including both synthetic and real scenarios. The synthetic set contains 18 easy scenarios (1-2 events and 2-6 individuals), and 10 hard scenarios (multiple events and 8 – 10 individuals). FM Synthetic dataset aims to capture group events without a complex object representation. Individuals are represented with simple colored circles as shown in Figure 19.

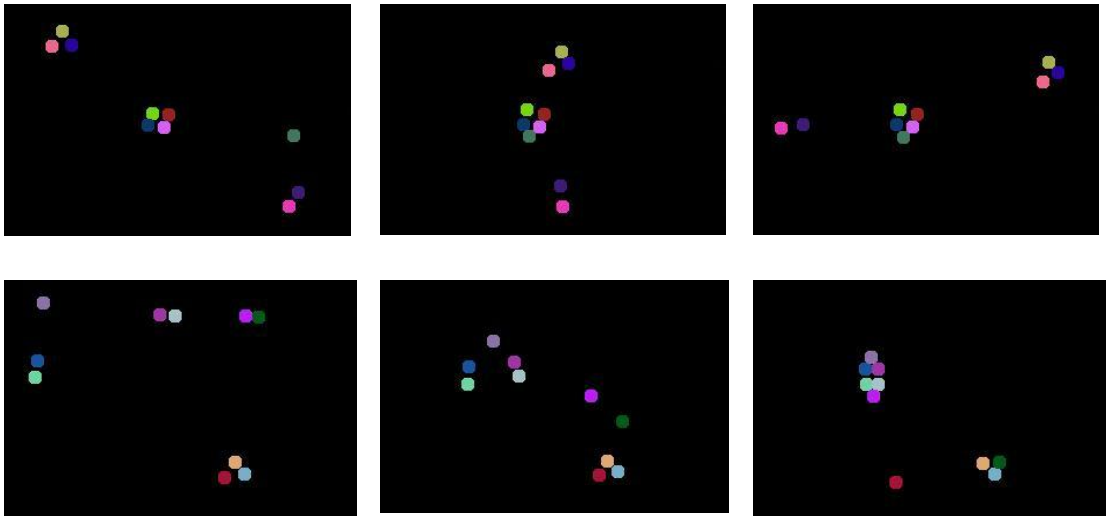


**Figure 19: FM Synthetic dataset [7, 43] object representation**

The FM dataset [7, 43] real set consists of outdoor scenes where individuals meet and has a total of 15 sequences varying from 30 seconds to 1.5 minutes. The real scenarios contain 3-11 individuals with multiple events shown in Figure 20.

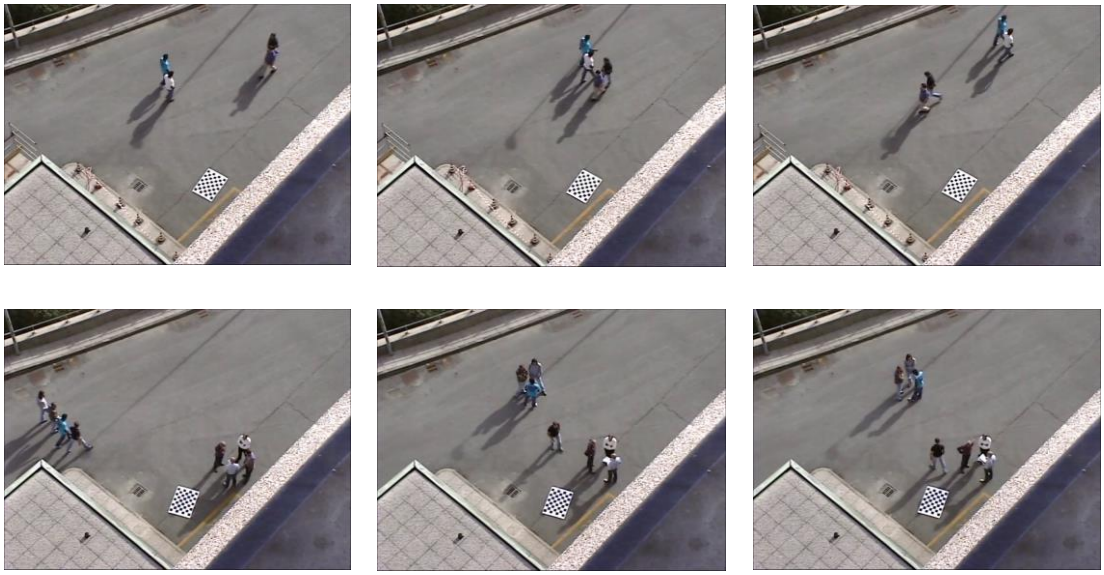


**Figure 20: FM Real Dataset [7, 43] with multiple events**

In FM synthetic and real dataset [7, 43], there are 3 synthetic and 2 queue sequences respectively which are excluded from our experiments since we focus on self-organizing groups. However, queue is a type of circumstantial group which is formed by some external forces [6].

In addition to FM dataset [7, 43], we conduct our tests on BIWI dataset [58] and PETS 2009 dataset [77] in order to assess the performance of proposed method. These datasets also have outdoor scenarios but do not contain self-organizing groups. Grouping events (merge and split) rarely occurs due to lack of suitable dataset and since BIWI dataset [58] is used in [7, 43] and Pets 2009 is tested with real person detector and its scenarios contain dense groups we conduct our experiments with these datasets.

## 5.2 Evaluation Method

As mentioned in Chapter 4, our CIGT framework is designed to track both individuals and groups. Therefore, we provide separate evaluation metrics for both individual and group tracking.

In individual tracking evaluation, we use False Positive (FP) and False Negative (FN) rates [79], ID-switch (IDS) [80], Mean Square Error (MSE) and Rooted Mean Square Error (RMSE) of the estimated positions and their standard deviation, MOTP [26] and MSE of overlapping area ratio between ground truth data and tracking result. Table 3 shows the individual tracking metrics used in evaluation of proposed tracking framework.

Although Overlap-Ratio denotes how tracking result overlap with the ground truth, MSE (px) and RMSE (px) denote the positional errors. These metrics are calculated as follows:

$$MSE = \frac{(x_i - x_{GT})^2 + (y_i - y_{GT})^2}{n} \tag{5.1}$$

$$RMSE = \sqrt{MSE} \tag{5.2}$$

$$Overlap_{ratio} = \frac{R_{GT} \cap R_{TR}}{R_{GT} \cup R_{TR}} \tag{5.3}$$

$$FP = \frac{n_{FP}}{n_{GT}} \tag{5.4}$$

$$FN = \frac{n_{FN}}{n_{GT}} \tag{5.5}$$

$$Re - Init = \frac{n_{reinit}}{n_{GT}} \tag{5.6}$$

where $n$ is the number of frames used in the evaluation, $(x_i, y_i)$ is the center of mass of tracker output, $(x_{GT}, y_{GT})$ is the center of mass of Ground Truth data, and $R_{GT}$ and $R_{TR}$ are the bounding rectangle area of individuals in ground truth and tracking result, respectively. The $n_{FP}$ is total number of false positives, $n_{FN}$ is the total number of false negatives, $n_{GT}$ is the total number of individual in ground truth and, $n_{reinit}$ is the total number of re-initialization from detection responses.

Before starting to explain individual and group tracking metrics, we provide the definitions of some tracking terms. Object is labelled as false positive if tracker finds an object which is not associated to ground truth. However, object is labelled as false negative if object exits in ground truth but tracker cannot find it. Tracker may drift the tracked object in time and rei-nitialize this object with detection responses. In this case, we increment re-initialization count and this metric shows us how well tracker models the motion pattern of object.

**Table 3: Individual Tracking Evaluation Metrics**

| Metric | Explanation | Desired Value |
|---|---|---|
| **1 – FP** | Percentage of individuals not identified as a False positive over total number of tracked objects. | Higher |
| **1 – FN** | Percentage of individuals not identified as a False negative over total number of tracked objects. | Higher |
| **Re-Init** | The average re-initialization rate per track. | Lower |
| **ID Switch** | The count of tracking result identified with a different ID in the previous frame. | Lower |
| **MSE (px)** | Square of positional error of tracking result in pixels. | Lower |
| **Overlap-Ratio** | Total overlapping ratio between tracking result and ground truth. | Higher |
| **RMSE (px)** | Positional error of tracking result in pixels. | Lower |

In our experiments, we evaluate groups in two ways: Detection and Tracking. For group detection, we use the Group Detection Success Rate (GDSR) [7, 43], False Positive (FP) and False Negative (FN) rates [79]. In group evaluation, we use convex hull covering all individuals in group. Therefore, in all group evaluation metrics, intersection operations are performed on the intersections among convex hulls of groups. Table 4 shows the group detection metrics and their explanation.

**Table 4: Group Detection Metrics**

| Metric | Explanation | Desired Value |
|---|---|---|
| **1 – FP** | Percentage of groups not identified as a False positive over total number of tracked objects. | Higher |
| **1 – FN** | Percentage of groups not identified as a False negative over total number of tracked objects. | Higher |
| **GDSR** | Number of time the groups is detected correctly. Group is assumed to be detected correctly if at least 60% of its members are detected. | Higher |

In group detection, FP and FN metrics are calculated by using equation 5.4 and 5.5, respectively. GDSR is calculated as follows:

$$GDSR = \frac{\sum_t \sum_i GD(g_t^i)}{\sum_t g_t}, where\ GD(g) = \begin{cases} 1, {}^{n_d}/_{n_g} \geq 0.6 \\ 0, otherwise \end{cases} \quad (5.7)$$

where $n_d$ is the detected group members by tracker, $n_g$ is the total number of group members, $g_t^i$ is the group *i* at frame *t*, $g_t$ is the number of groups at frame *t*.

In tracking of group, we use the Multi-Object Tracking Precision (MOTP) and Accuracy (MOTA) [81] metrics shown in Table 5.

**Table 5: Group Tracking Metrics**

| Metric | Explanation | Desired Value |
|---|---|---|
| **MOTP** | Total error in estimated position for matched object-hypothesis pairs for all frames. | Lower |
| **MOTA** | Percentage of tracking result without errors over total number of matched objects. | Higher |

MOTP evaluates tracker by means of precise object position and estimates the total position error over all matched frames as follows:

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \quad (5.8)$$

where $d_t^i$ is the distance between Ground Truth position and estimated position for object and $c_t$ is the number of matched frames for frame t. MOTA evaluates the accuracy of tracker by means of number of misses, false positives and mismatches as follows:

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \tag{5.9}$$

where $m_t$ is the number of misses, $fp_t$ is the number of false positives, $mme_t$ is the number of mismatches and $g_t$ is the number of objects in ground truth data for frame *t*.

In our experiments, we calculate MOTP in meters by using Homography matrix. In FM dataset [7, 43] and BIWI dataset [58] we convert each detection position as follows:

$$[y_m \quad x_m \quad z_m]^T = H \cdot [y \quad x \quad 1]^T \tag{5.10}$$

where (x, y) is the pixel coordinate of tracked object, $x_m = \frac{x_m}{z_m}$, $y_m = \frac{y_m}{z_m}$ are the corresponding coordinate in meters.

For PETS 2009 dataset [77], no Homography matrix was provided. However, we calculate this matrix by using the camera parameters for View 001.

$$H = \begin{bmatrix} f \cdot sx/dpx & 0 & Cx \\ 0 & f/dpy & Cy \\ 0 & 0 & 1 \end{bmatrix} \tag{5.11}$$

where $(f, sx, dpx, dpy)$ composes 2x1 focal vector, $(Cx, Cy)$ is the principal point. In our experiment, camera parameters for PETS 2009 dataset [77] are shown in Table 6.

**Table 6: Camera Parameters for View 001 in PETS 2009 Dataset [77]**

| Camera parameter | Value |
|---|---|
| ***f*** | 5.5549183034 |
| ***sx*** | 1.0937855397 |
| ***dpx*** | 0.0051273271277 |
| ***dpy*** | 0.00465 |
| ***Cx*** | 324.22149053 |
| ***Cy*** | 282.56650051 |

## 5.3 Multi Object Tracking Evaluation

In this section, we evaluate the performance of our proposed method against those of the state-of-art methods by means of metrics explained in section 5.2. In addition, we evaluate the each video scenario separately.

Datasets can be classified into synthetic scenarios and real scenarios since synthetic and real scenarios have different visual representation of object. Therefore, we divide out multi object tracking evaluation into two parts.

### 5.3.1 Results on synthetic scenarios

In this part, we evaluate CIGT framework on FM synthetic data by comparing with DP2-JIGT [43] and DEEPER-JIGT [7]. Table 7 shows the results on FM synthetic dataset.

**Table 7: Results on the FM synthetic dataset excluding queue sequences**

| | **MSE [px] (std)** | **1-FP** | **1-FN** | **GDSR** | **MOTP [px]** | **MOTA** |
|---|---|---|---|---|---|---|
| **CIGT** | 3.43 (3.95) | **94.79%** | 90.37% | **90.26%** | **2.37** | **86.36%** |
| **DP2-JIGT [43]** | **1.75** (4.76) | 93.98% | **91.28%** | 86.91% | 16.72 | 71.57% |
| **DEEPER-JIGT [7]** | 2.28 (5.08) | 93.12% | 81.01% | 78.18% | 18.16 | 53.42% |

In Table 7, only the MSE (px) is an individual tracking metric while others are group evaluation metrics. CIGT framework uses both template detector and particle advection to model the motion and adjusts weights automatically according to the group density. In individual tracking, since template detector has a higher weight than particle advection, CIGT framework is affected by detection noises more than both DP2-JIGT [43] and DEEPER-JIGT [7] and has higher MSE compared to DP2-JIGT [43] and DEEPER-JIGT [7]. However, it outperforms DP2-JIGT and DEEPER-JIGT for group detection and tracking. Only DP2-JIGT [43] has slightly higher than our proposed CIGT framework since DP2-JIGT [43] models group in an online fashion. However, as a result of false positive elimination mechanism and evaluation of closeness and motion direction in multi-observation model, CIGT framework has slightly better performance for 1-FP and GDSR metrics. Also, as a result of strength of particle advection, CIGT framework is better to model the group motion and has lower MOTP than other works in state of art. Because of two-phase association and false positive

elimination, CIGT framework has better ID switch, false positive and false negative results on individual tracking shown in Table 8. As a result, a CIGT framework has better MOTA compared to DP2-JIGT [43] and DEEPER-JIGT [7].

According to Table 8, CIGT framework is not affected with respect to different FM Dataset [7, 43] synthetic scenarios for individual tracking. Because of false positive elimination mechanism in CIGT framework, there is no significant difference between 1-FP metric of different scenarios compared to 1-FN metric. Due to object representation of FM Dataset [7, 43] synthetic (one-colored circle shown in Figure 19), region covariance and HOG features are not effective on appearance model. Objects in synthetic scenarios are similar to each other with aspect of shape and texture. Therefore, color histogram is dominant feature to discriminate object from each other for FM Dataset [7, 43] synthetic scenarios. As a result of this, ID switch count increases in some scenarios where objects are very close to each other. The other interesting metric is the standard deviation of MSE and RSME. All standard deviation metrics are close to each other for all FM Dataset [7, 43] synthetic scenarios. Recall that template match is more effective than particle advection in motion model for individual tracking and this causes that individual tracking results are more sensitive to detection noise. Since tracking results for all individuals are affected by detection noise at the close rate, we obtain standard deviation metric that is close to each other for all individual tracking result.

Group detection and tracking results for all FM Dataset [7, 43] synthetic scenarios are shown in Table 9. As a result of false positive elimination mechanism in CIGT framework, 1-FP metric is less affected compared to 1-FN metric. In some scenarios (Hard4 and Merge5 in Table 9), 1-FN metric decreases to below 80%. Ground truth data is generated according to mostly distance between objects. However, in CIGT framework, we consider not only distance but also direction angle between objects. Therefore, in some cases, our framework detects split event and form new groups while there is no split event and group change in ground truth data. As a result of this, our 1-FN metric decreases in some cases. Also, MOTA and MOTP (px) are also very good as a result of individual tracking performance.

**Table 8: CIGT Framework detailed result on FM Dataset [7, 43] synthetic scenarios for individual tracking**

| Video | 1-FP | 1-FN | Re-Init | IDS | Overlap-Ratio | MSE | MSE std | RMSE | RMSE std |
|---|---|---|---|---|---|---|---|---|---|
| **Hard1** | 93.78% | 92.17% | 0.06% | 30 | 0.69 | 3.34 | 3.80 | 1.56 | 0.96 |
| **Hard2** | 95.00% | 92.64% | 0.07% | 10 | 0.69 | 3.55 | 4.09 | 1.61 | 0.98 |
| **Hard3** | 93.19% | 93.56% | 0.06% | 9 | 0.69 | 3.52 | 4.04 | 1.59 | 0.99 |
| **Hard4** | 94.80% | 93.30% | 0.05% | 9 | 0.69 | 3.54 | 4.00 | 1.61 | 0.97 |
| **Hard5** | 94.45% | 93.75% | 0.05% | 15 | 0.69 | 3.48 | 3.99 | 1.59 | 0.97 |
| **Hard6** | 94.90% | 94.75% | 0.05% | 8 | 0.69 | 3.37 | 3.86 | 1.57 | 0.95 |
| **Hard7** | 94.95% | 87.05% | 0.05% | 31 | 0.69 | 3.44 | 3.93 | 1.59 | 0.96 |
| **Hard8** | 93.13% | 92.19% | 0.06% | 20 | 0.69 | 3.42 | 4.01 | 1.57 | 0.98 |
| **Hard9** | 94.00% | 92.10% | 0.05% | 30 | 0.69 | 3.49 | 4.16 | 1.58 | 1.00 |
| **Hard10** | 93.15% | 91.25% | 0.05% | 35 | 0.69 | 3.41 | 3.92 | 1.57 | 0.97 |
| **Merge1** | 93.38% | 93.75% | 0.00% | 4 | 0.70 | 3.19 | 3.66 | 1.54 | 0.91 |
| **Merge2** | 95.75% | 94.00% | 0.00% | 21 | 0.69 | 3.28 | 3.65 | 1.57 | 0.91 |
| **Merge3** | 93.25% | 94.38% | 0.00% | 4 | 0.69 | 3.41 | 3.94 | 1.58 | 0.96 |
| **Merge4** | 93.88% | 94.50% | 0.00% | 2 | 0.67 | 3.55 | 4.00 | 1.62 | 0.97 |
| **Merge5** | 93.60% | 92.00% | 0.00% | 30 | 0.69 | 3.33 | 3.75 | 1.57 | 0.94 |
| **Opposite1** | 94.38% | 95.25% | 0.00% | 0 | 0.68 | 3.62 | 4.08 | 1.64 | 0.97 |
| **Opposite2** | 93.63% | 94.50% | 0.00% | 0 | 0.69 | 3.50 | 4.10 | 1.59 | 0.99 |
| **Opposite3** | 94.92% | 95.50% | 0.00% | 0 | 0.70 | 3.23 | 3.61 | 1.55 | 0.91 |
| **Opposite4** | 93.00% | 94.75% | 0.00% | 0 | 0.68 | 3.63 | 4.06 | 1.63 | 0.98 |
| **Opposite5** | 93.67% | 94.83% | 0.00% | 0 | 0.69 | 3.43 | 4.16 | 1.56 | 0.99 |
| **Split1** | 95.13% | 96.00% | 0.00% | 0 | 0.69 | 3.22 | 3.74 | 1.53 | 0.93 |
| **Split2** | 94.00% | 93.63% | 0.00% | 16 | 0.69 | 3.52 | 4.19 | 1.60 | 0.99 |
| **Split3** | 94.38% | 92.25% | 0.00% | 28 | 0.69 | 3.51 | 4.16 | 1.60 | 0.97 |
| **Split4** | 91.90% | 88.70% | 0.00% | 20 | 0.70 | 3.45 | 4.10 | 1.58 | 0.98 |
| **Split5** | 94.00% | 89.60% | 0.00% | 33 | 0.69 | 3.34 | 3.83 | 1.56 | 0.95 |
| **Average** | **94.01%** | **94.11%** | **0.02%** | **14.20** | **0.69** | **3.43** | **3.95** | **1.58** | **0.96** |

**Table 9: CIGT Framework detailed result on FM Dataset [7, 43] synthetic scenarios for group detection and tracking**

| Video | 1-FP | 1-FN | MOTP (px) | MOTA | GDSR |
|---|---|---|---|---|---|
| **Hard1** | 88.73% | 84.92% | 3.14 | 85.94% | 84.92% |
| **Hard2** | 93.03% | 91.04% | 2.29 | 87.64% | 91.04% |
| **Hard3** | 87.28% | 84.05% | 3.32 | 86.75% | 84.05% |
| **Hard4** | 98.33% | 97.50% | 2.40 | 88.10% | 97.50% |
| **Hard5** | 86.53% | 75.16% | 2.26 | 88.20% | 74.68% |
| **Hard6** | 93.66% | 95.15% | 2.35 | 89.65% | 94.96% |
| **Hard7** | 96.36% | 91.06% | 2.46 | 82.00% | 90.89% |
| **Hard8** | 96.75% | 96.25% | 2.54 | 85.31% | 96.25% |
| **Hard9** | 89.09% | 82.21% | 2.51 | 86.10% | 82.21% |
| **Hard10** | 92.23% | 87.45% | 2.76 | 84.40% | 87.45% |
| **Merge1** | 97.79% | 97.06% | 1.97 | 87.13% | 97.06% |
| **Merge2** | 92.51% | 80.52% | 2.18 | 89.75% | 80.52% |
| **Merge3** | 98.00% | 97.50% | 2.40 | 87.63% | 97.50% |
| **Merge4** | 93.52% | 98.98% | 2.53 | 88.38% | 98.98% |
| **Merge5** | 94.31% | 78.93% | 2.88 | 85.60% | 78.26% |
| **Opposite1** | 99.00% | 98.50% | 2.26 | 89.63% | 98.50% |
| **Opposite2** | 99.75% | 97.00% | 2.41 | 88.13% | 97.00% |
| **Opposite3** | 100.00% | 99.50% | 2.00 | 90.42% | 99.00% |
| **Opposite4** | NA | NA | NA | 87.75% | NA |
| **Opposite5** | 100.00% | 99.00% | 2.10 | 88.50% | 99.00% |
| **Split1** | 96.77% | 97.85% | 2.00 | 91.13% | 97.13% |
| **Split2** | 97.50% | 97.00% | 2.02 | 87.63% | 97.00% |
| **Split3** | 93.53% | 86.47% | 1.93 | 86.63% | 86.47% |
| **Split4** | 92.75% | 92.75% | 2.12 | 80.60% | 92.27% |
| **Split5** | 95.70% | 90.73% | 2.34 | 83.60% | 90.73% |
| **Average** | **94.71%** | **91.52%** | **2.38** | **87.06** | **91.39** |

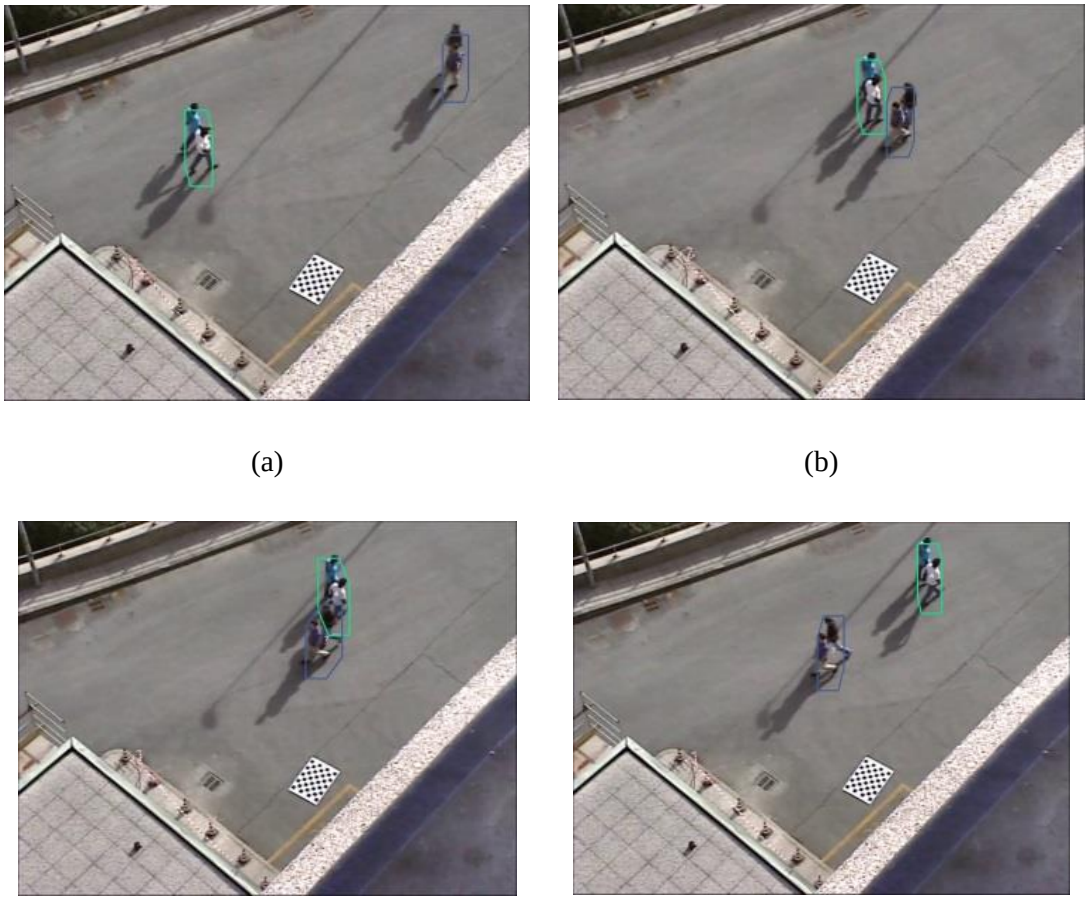
### 5.3.2 Results on real scenarios

Evaluation on real datasets is performed on three different sets: FM [7, 43], BIWI [58] and Pets 2009 [77]. All three sets have different properties and challenges. However, most considerable dataset is the real part of FM dataset [7, 43] since it provides so many group events in scenarios. Table 10 shows the group evaluation on real part of FM Dataset [7, 43].

**Table 10: Results on the real FM dataset excluding queue sequences for group detection and tracking**

| | **1-FP** | **1-FN** | **GDSR** | **MOTP [m]** | **MOTA** |
|---|---|---|---|---|---|
| **CIGT** | 97.40% | 95.81% | **95.81%** | **0.07** | **94.79%** |
| **DP2-JIGT [43]** | **97.81%** | **97.54%** | 94.65% | 0.92 | 73.85% |
| **DEEPER-JIGT [7]** | 95.72% | 89.99% | 85.78% | 0.87 | 65.18% |

In group detection result, DP2-JIGT [43] has slightly better performance than CIGT framework with respect to 1-FP and 1-FN metrics because of online group modelling in DP2-JIGT [43]. However, CIGT framework has slightly better performance than DP2-JIGT [43] since CIGT framework evaluates not only distance but also movement direction of individuals in its multi-observation model. This feature of multi-observation model provides us to identify different groups with different motion direction and to prevent merging with different groups as shown in Figure 21. In this scenario, two different groups approach to each other. However, merge event is not occurred even if these groups are very close to each other shown in Figure 21 (b) and (c) due to different movement direction.



(a) (b)

(c) (d)

**Figure 21: CIGT Framework evaluation of different groups with different movement direction**

As shown in Table 10, CIGT framework outperforms DP2-JIGT [43] and DEEPER-JIGT [43] for group tracking statistics. Because of the particle advection used in motion model and multi observation model, CIGT framework is better positioned the groups and gets lower MOTP [m] errors and higher MOTA metric compared to DP2-JIGT [43] and DEEPER-JIGT [43]. Individual tracking results on real part of FM dataset [7, 43] is shown in Table 11.

**Table 11: Results on the real part of FM dataset [7, 43] excluding queue sequences for individual tracking**

| | 1-FP | 1-FN | Overlap-Ratio | Re-Init | ID |
|---|---|---|---|---|---|
| **CIGT** | **96.60%** | **98.52%** | **0.79** | **0.1%** | **14** |
| **DP2-JIGT [43]** | 81.25% | 78.11% | 0.71 | 3.3% | 156 |
| **DEEPER-JIGT [7]** | 95.72% | 89.99% | 0.71 | 3.2% | 148 |

CIGT framework uses discriminative appearance model [5] to describe the individuals and performs two-phase association to identify them by using features obtained from discriminative appearance model [5]. This provides CIGT to identify individuals more precisely. As a result of this, CIGT framework outperforms DP2-JIGT [43] and DEEPER-JIGT [7] by means of ID switch, 1-FP and 1-FN metrics. Also, as a result of particle advection mechanism in CIGT framework for motion modeling, it is more precisely positioned individuals, CIGT tracker does not need to be reinitialized as often as DP2-JIGT [43] and DEEPER-JIGT [7] and as result of this, it has better Overlap-Ratio and Re-Init results.

Other than comparing our proposed CIGT framework with works in state-of-art, we also evaluate real scenario individually on FM Dataset [3, 19]. Table 12 shows the individual results on real scenarios on FM Dataset [3, 19].

**Table 12: CIGT Framework detailed result on FM Dataset [3, 19] real scenarios for individual tracking**

| Video | 1-FP | 1-FN | Re-Init | IDS | Overlap-Ratio | MSE [px] | MSE std | RMSE | RMSE std |
|---|---|---|---|---|---|---|---|---|---|
| **S01** | 96.72% | 98.22% | 0.09% | 7 | 0.79 | 41.26 | 39.81 | 5.73 | 2.90 |
| **S02** | 98.34% | 98.66% | 0.08% | 8 | 0.79 | 41.11 | 38.56 | 5.72 | 2.90 |
| **S03** | 96.93% | 99.32% | 0.21% | 1 | 0.80 | 37.43 | 35.96 | 5.45 | 2.78 |
| **S04** | 96.11% | 99.10% | 0.00% | 2 | 0.78 | 41.02 | 38.00 | 5.74 | 2.85 |
| **S05** | 96.71% | 98.15% | 0.04% | 15 | 0.79 | 37.37 | 35.00 | 5.45 | 2.76 |
| **S06** | 98.66% | 99.00% | 0.00% | 2 | 0.79 | 39.96 | 38.18 | 5.62 | 2.90 |
| **S07** | 98.57% | 98.57% | 0.06% | 34 | 0.79 | 39.40 | 39.03 | 5.58 | 2.88 |
| **S08** | 90.04% | 98.44% | 0.20% | 7 | 0.78 | 43.37 | 40.29 | 5.88 | 2.96 |
| **S09** | 97.12% | 98.54% | 0.11% | 5 | 0.79 | 41.00 | 41.14 | 5.70 | 2.93 |
| **S10** | 98.23% | 96.31% | 0.10% | 23 | 0.79 | 40.28 | 39.56 | 5.63 | 2.93 |
| **S11** | 91.40% | 99.02% | 0.12% | 0 | 0.79 | 39.17 | 38.00 | 5.58 | 2.84 |
| **S14** | 98.11% | 98.76% | 0.04% | 32 | 0.79 | 40.45 | 38.45 | 5.67 | 2.87 |
| **S15** | 98.88% | 94.27% | 0.01% | 36 | 0.79 | 39.83 | 37.74 | 5.62 | 2.86 |
| **Average** | **96.60%** | **98.52%** | **0.08%** | **13.23** | **0.79** | **40.13** | **38.44** | **5.64** | **2.87** |

According to Table 12, all metrics for individual tracking are consistent with each on different scenarios. That shows us CIGT framework is not so much affected with respect to different scenarios. Only ID switch metrics vary according to scenarios compared to other metrics because of following reasons:

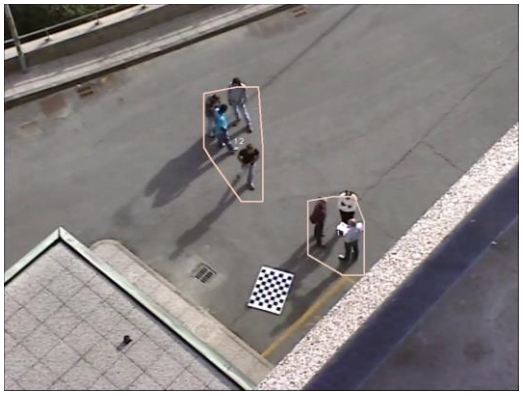
(i) These scenarios have longer durations compared to others

(ii) ID switch metric is obtained by consecutive frames. However, CIGT framework evaluates all video sequences and if there is tracked object with wrong ID, it corrects the ID for tracked object. For example, we have tracked object with ID = 5. On next frame, CIGT does not associate object with ID = 5 and gives ID = 6 to this object. However, few frames later, CIGT set object ID = 5 to this object. In this case, we count ID switch as 2 but in reality, CIGT switched back to the original (correct) ID for this object.

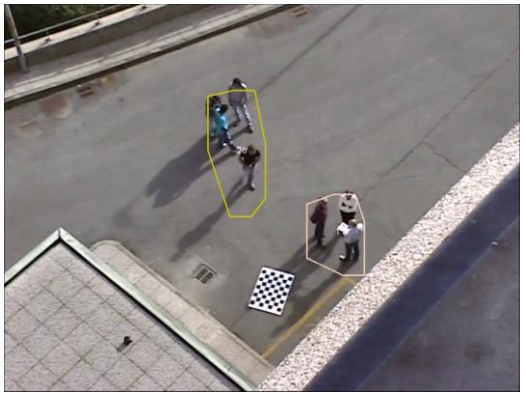
(iii) Individuals with similar appearances are close to each other shown in Figure 22.

(iv) Two detections are obtained with same tracked object shown in Figure 23. Although groups in Figure 23 (a) and (b) are the same, they are identified with different label. Because we get two detections for individuals shown in Figure 23 (d) and our false positive elimination mechanism cannot discard the one of the detection responses. As a result, our CIGT framework evaluates the merge events and set different ID for group shown in Figure 23 (b).

**Figure 22: ID switch for individual tracking since similar individuals are too close**

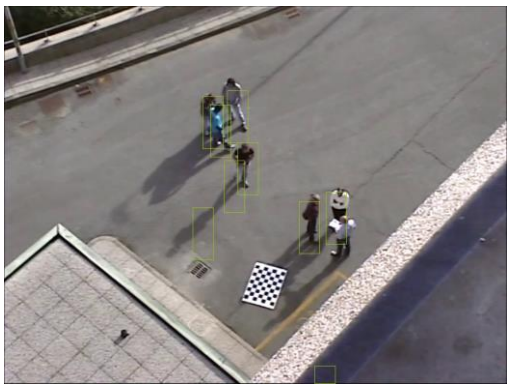


(a) CIGT Tracking result



(b) Same group with (a) but different ID

(c) Detection responses in (a)



(d) Detection responses in (b)

**Figure 23: ID switch for individual tracking due to two detections in single individual**

Group detection and tracking result with respect to different real scenarios on FM dataset [7, 43] are shown in Table 13. There is no significant difference on metrics obtained from different scenarios. As a result of multi-observation model and false positive elimination mechanism, it has good performance for group detection metrics. Also, because of particle advection and two-phase association, CIGT tracks group with small MOTP error and high MOTA for group tracking.

**Table 13: CIGT Framework detailed result on FM Dataset [3, 19] real scenarios for group detection and tracking**

| Video | 1-FP | 1-FN | MOTP [px] | MOTP [m] | MOTA | GDSR |
|---|---|---|---|---|---|---|
| **S01** | 93.84% | 90.06% | 4.04 | 0.06 | 94.94% | 90.06% |
| **S02** | 98.43% | 98.43% | 3.71 | 0.05 | 97.00% | 98.43% |
| **S03** | 99.44% | 99.44% | 3.83 | 0.06 | 96.26% | 99.44% |
| **S04** | 97.30% | 88.87% | 4.62 | 0.07 | 95.21% | 88.87% |
| **S05** | 97.31% | 97.82% | 4.35 | 0.06 | 94.86% | 97.82% |
| **S06** | 100.00% | 100.00% | 4.71 | 0.07 | 97.67% | 100.00% |
| **S07** | 98.43% | 97.40% | 5.33 | 0.08 | 97.15% | 97.40% |
| **S08** | 95.64% | 96.68% | 6.05 | 0.08 | 88.48% | 96.68% |
| **S09** | 99.03% | 99.45% | 4.11 | 0.06 | 95.67% | 99.45% |
| **S10** | 97.43% | 95.77% | 3.74 | 0.06 | 94.55% | 95.77% |
| **S11** | 97.70% | 93.11% | 6.42 | 0.09 | 90.42% | 93.11% |
| **S14** | 98.27% | 98.88% | 4.62 | 0.07 | 96.87% | 98.88% |
| **S15** | 93.32% | 89.62% | 6.71 | 0.09 | 93.15% | 89.62% |
| **Average** | **97.40%** | **95.81%** | **4.79** | **0.07** | **94.79%** | **95.81%** |

The BIWI dataset [58] consists of two video sequences which individuals generally walk in one direction and is very poor for group events: merge and split. Due to lack of dataset and since DP2-JIGT [43] and DEEPER-JIGT [7] are evaluated with this dataset, we conduct experiments on BIWI dataset [58] and results are shown in Table 14.

**Table 14: Results on the BIWI dataset. Columns (1-3) for group detection, columns (4-5) group tracking**

| | 1-FP | 1-FN | GDSR | MOTP [m] | MOTA |
|---|---|---|---|---|---|
| **CIGT** | **91.49%** | 56.16% | **54.57%** | **0.31** | **29.66%** |
| **DP2-JIGT [43]** | 37.66% | **89.43%** | 51.86% | 0.47 | 22.94% |
| **DEEPER-JIGT** | 53.77% | 78.00% | 53.59% | 0.44 | 29.43% |

As a result of false positive elimination in CIGT framework, it outperforms DP2-JIGT [43] and DEEPER-JIGT [7] for false positive detection. However, due to bad illumination and too

small detection, discriminative appearance model [5] cannot describe objects very well and foreground objects are not extracted all the time. These factors cause increase of false negative ratio. Because of multi-observation model in CIGT, GDSR metric is slightly higher than DP2-JIGT [43] and DEEPER-JIGT [7]. Also, CIGT framework has slightly higher than DP2-JIGT [43] and DEEPER-JIGT [7] according to MOTP [m] and MOTA since particle advection in CIGT framework provide better group motion modeling and false positive detections are mostly eliminated by CIGT framework.

In addition to comparing our proposed CIGT framework with works in state-of-art, we also provide detailed result of individual and group tracking shown in Table 15 and Table 16, respectively.

**Table 15: CIGT Framework detailed result on BIWI Dataset [58] for individual tracking**

| Video | 1-FP | 1-FN | Re-Init | IDS | Overlap-Ratio | RMSE | RMSE std |
|---|---|---|---|---|---|---|---|
| **eth** | 99.42% | 71.68% | 2.03% | 1,768 | 0.48 | 6.53 | 3.30 |
| **hotel** | 99.69% | 49.73% | 4.32% | 2,707 | 0.41 | 13.01 | 4.23 |
| **Average** | **99.56%** | **60.71%** | **3.18%** | **2237.50** | **0.45** | **9.77** | **3.76** |

In individual tracking on BIWI Dataset [58], CIGT framework eliminates %99.56 of false positive detections because of its false positive elimination mechanism. However, due to bad illumination and noise, individuals are not identified very well and as a result of this, ID switch count increases. Also, detection noise causes high Overlap-Ratio, RMSE.

**Table 16: CIGT Framework detailed result on BIWI Dataset [10] for group detection and tracking**

| Video | 1-FP | 1-FN | MOTP [px] | MOTP [m] | MOTA | GDSR |
|---|---|---|---|---|---|---|
| **eth** | 87.50% | 61.83% | 8.78% | 0.39 | 51.25% | 59.63% |
| **hotel** | 95.49% | 50.49% | 13.27% | 0.24 | 8.07% | 49.51% |
| **Average** | **91.49%** | **56.16%** | **11.03%** | **0.31** | **29.66%** | **54.57%** |

According to group detection and tracking results shown in Table 16, CIGT framework has low 1-FN metric since BIWI Dataset [58] has poor merge and split events and is not captured for self-organizing groups. Due to ID switch and 1-FN in individual tracking results, CIGT has lower MOTA compared to results on FM Dataset [7, 43]. Also, bad illumination and noise cause the degradation on particle advection in motion model and consequently, MOTP metric get lower compared to MOTP on FM Dataset [7, 43].

The final challenging dataset that we used in our experiments is the PETS 2009 dataset [77]. Scenarios of PETS 2009 dataset [77] includes denser groups compared to other datasets and unlike other experiment, we use real person detector [78] to evaluate tracker performance. Individual and group tracking results on PETS 2009 [77] are shown in Table 17 and Table 18, respectively.

As a result of false positive elimination mechanism, CIGT has high 1-FP metric. However, due to camera view, individuals get bigger or smaller depending on their movement direction and this causes different appearance representation for same individuals. This causes the increase on ID switch and decrease on MOTA and GDSR.

**Table 17: CIGT Framework detailed result on PETS 2009 Dataset [77] for individual tracking**

| Video | 1-FP | 1-FN | Re-Init | IDS | Overlap-Ratio | RMSE [px] | RMSE std |
|---|---|---|---|---|---|---|---|
| **S1L1-1** | 92.15% | 68.31% | 0.64% | 417 | 0.62 | 9.86 | 10.63 |
| **S1L1-2** | 93.84% | 69.29% | 0.94% | 161 | 0.58 | 9.18 | 7.54 |
| **S1L2-1** | 97.34% | 39.50% | 0.76% | 331 | 0.45 | 14.40 | 10.46 |
| **S1L2-2** | 97.12% | 44.56% | 0.88% | 254 | 0.53 | 13.86 | 13.17 |
| **S2L1** | 86.56% | 96.30% | 0.15% | 72 | 0.63 | 5.80 | 5.27 |
| **S2L2** | 98.38% | 46.78% | 0.82% | 184 | 0.55 | 9.37 | 8.47 |
| **Average** | **94.23%** | **60.79%** | **0.70%** | **236.50** | **0.79** | **10.41** | **9.26** |

**Table 18: CIGT Framework detailed result on PETS 2009 Dataset [22] for group detection and tracking**

| | 1-FP | 1-FN | MOTP [m] | MOTP [px] | MOTA | GDSR |
|---|---|---|---|---|---|---|
| **S1L1-1** | 98.83% | 60.08% | 1.41 | 28.33 | 52.06% | 59.30% |
| **S1L1-2** | 96.64% | 60.62% | 1.30 | 30.04 | 58.94% | 60.62% |
| **S1L2-1** | 100.00% | 75.32% | 1.09 | 26.26 | 30.22% | 74.04% |
| **S1L2-2** | 93.49% | 83.26% | 1.40 | 28.98 | 35.27% | 81.86% |
| **S2L1** | 68.41% | 83.31% | 1.28 | 16.32 | 81.31% | 81.52% |
| **S2L2** | 91.30% | 66.67% | 0.54 | 14.65 | 40.95% | 66.43% |
| **Average** | **%91.44** | **%71.54** | **1.17** | **24.09** | **49.79** | **70.63** |

In summary, CIGT shows good performance compared to the other works in literature. Multi-observation model provides handling group events and forming and updating group state according to these events. Also, particle advection performs great ability to estimate group motion patterns.

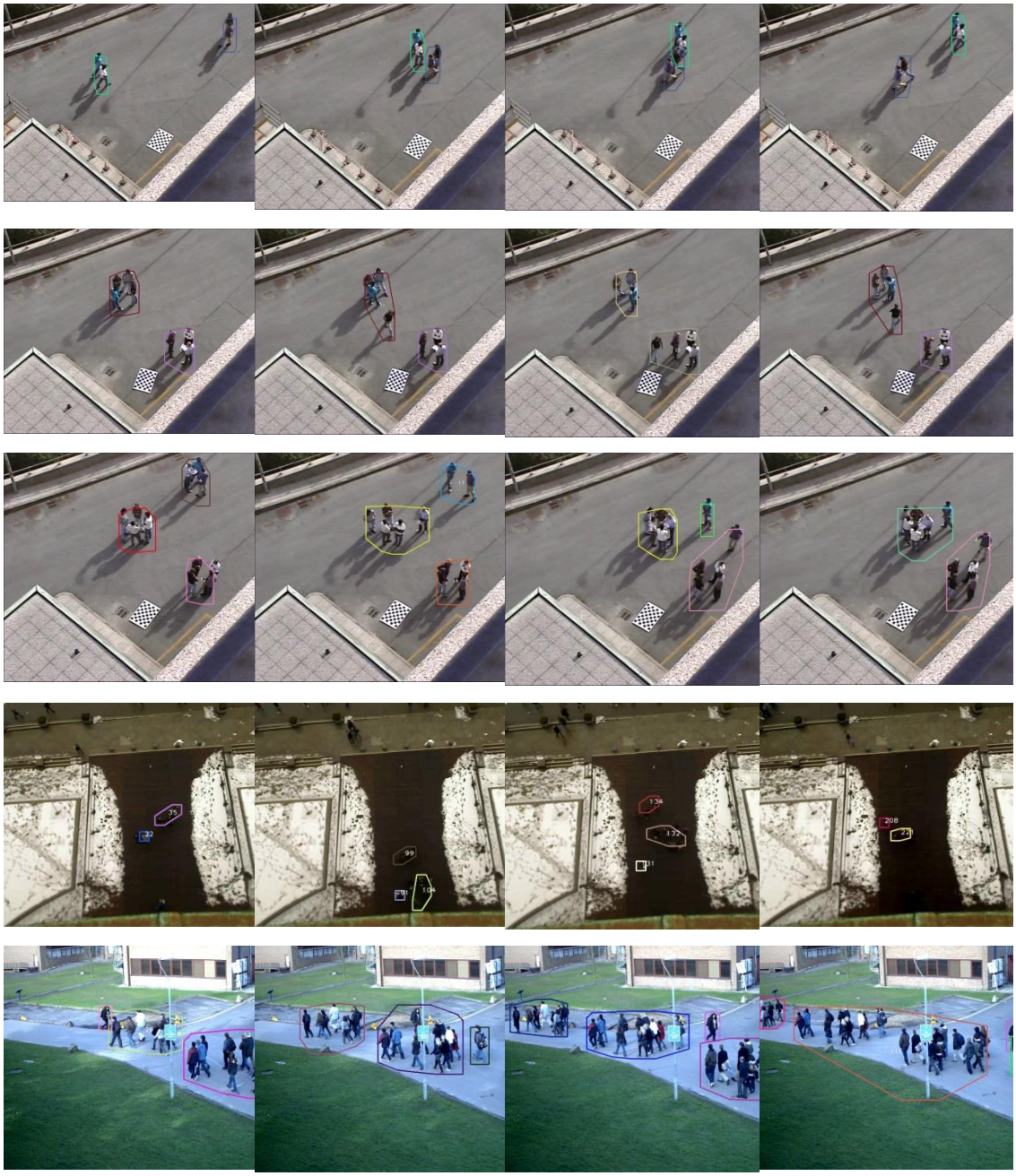


**Figure 24: Visual Results of CIGT framework on FM Dataset [7, 43], BIWI Dataset [58], PETS 2009 Dataset [22]**

## 5.4 Person Detection Evaluation

In this section, we provide analysis on the effect of person detection error on the performance of the proposed CIGT framework. Recall that we used the simulation of person detection by generating 20% of false positive and negative from ground truth data in previous experiments. In this experiment, we use this person detection simulation with different error ratios varying from 0.00 to 0.40 by increments of 0.05 and analyze statistics from this person detection results for both individual and group tracking. Person detection simulator puts only one false positive and one false negative in detection result in case that error ratio is 0.00.

Since our proposed framework is designed for both individuals and groups, we evaluate how much both individual and group tracking are affected.

### 5.4.1 Person Detection Effect on Individual Tracking

Individual tracking is evaluated by means of 1-FP, 1-FN, ID switch and Re-Init metrics. Figure 25 shows the person detection effect on 1-FP and 1-FN for individual tracking.

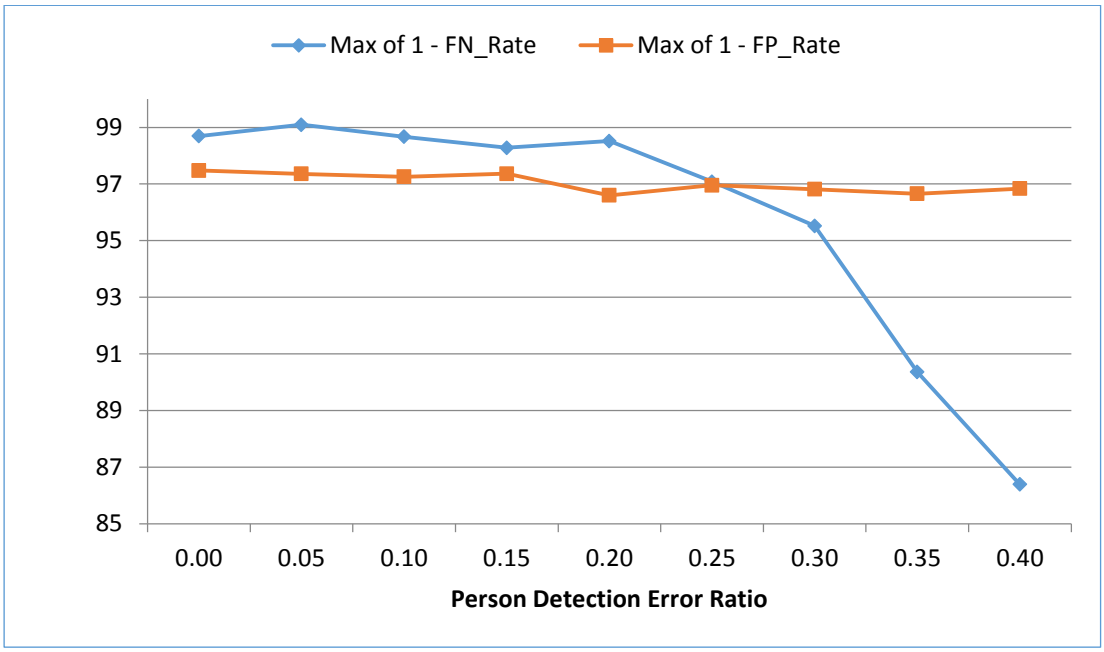



**Figure 25: Person detection effect on individual tracking by means of 1-FP and 1-FN metrics.**

Although 1-FN decreases after error ratio is greater than 0.25, there is no significant change in 1-FP metric since false positive elimination mechanism in CIGT framework mostly dis-

cards false positive from detection responses with hierarchical approach explained in 4.3. However, the decrease in 1-FN is not proportional to error ratio and still greater than 85% in case that error ratio is 0.40. The other important metric showing individual tracking success is the ID switch count and person detection effect on this metric is shown in Figure 26.

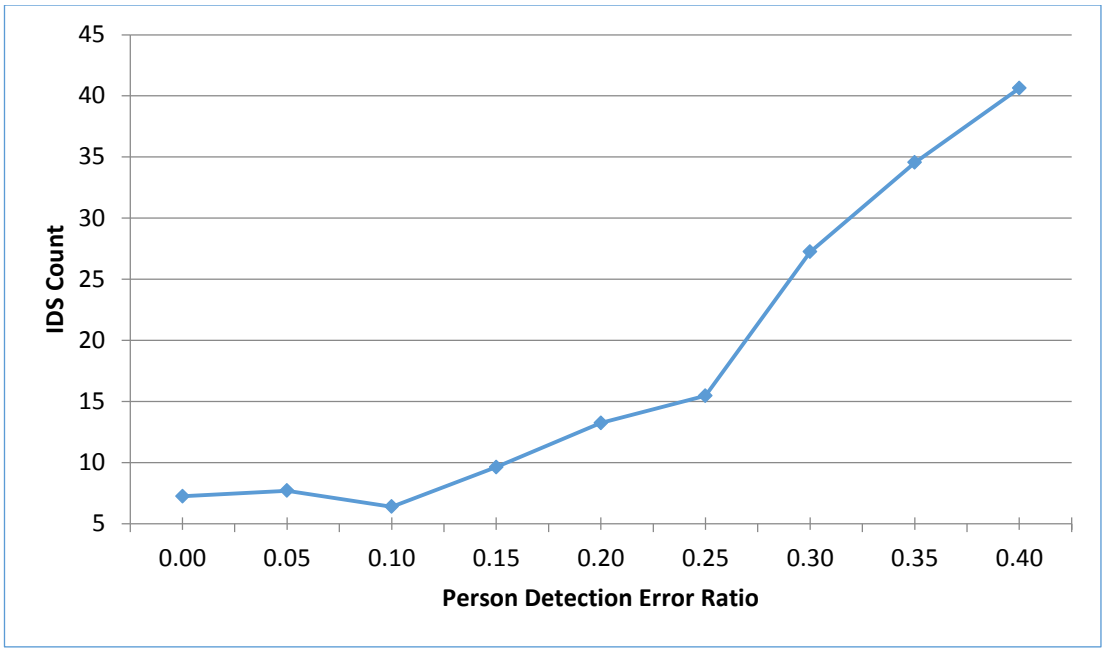



**Figure 26: Person detection effect on individual tracking by means of IDS metrics.**

As a result of the increase on false negative detection responses, tracked objects are not updated as often as lower error rates. As a result of this, low-level association rate is decreased and number of ID switches increase. Mid-level association can compensate low-level associations until error rate is 0.25 but after that ratio, ID switch count increases. However, ID switch count in CIGT is still lower than which similar tracker's in literature.

**Recall that Re-Init rate shows us how often tracker is reinitialized with detection responses due to its drift. This metric analyzes the motion model of proposed method and result is shown in**

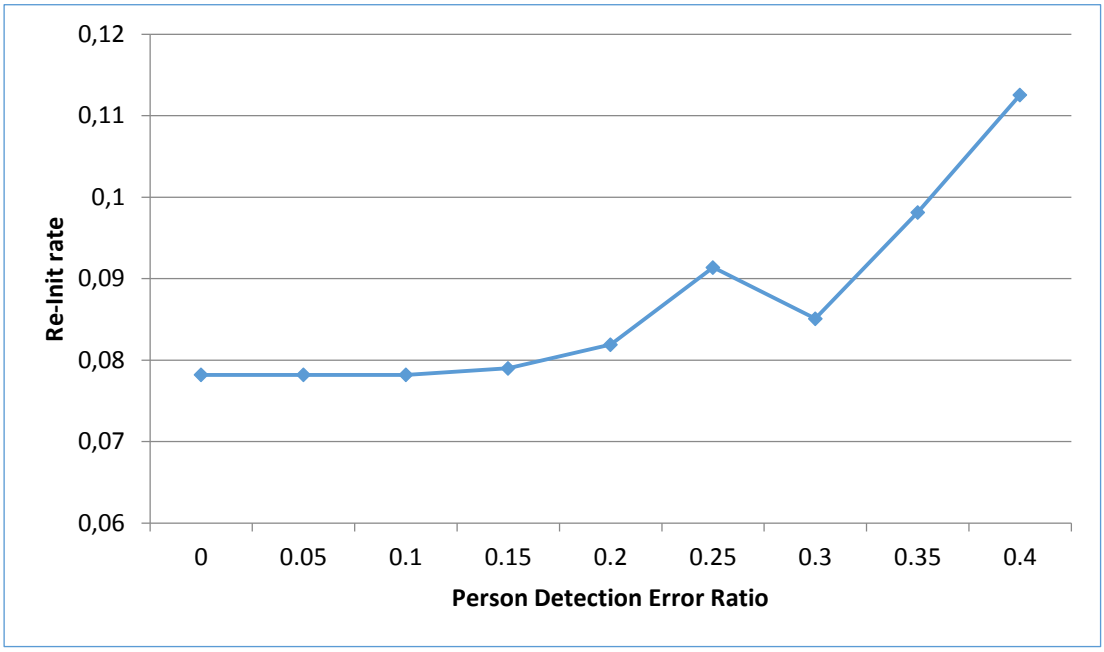



Figure 27. There is no significant change on Re-Init rate with increasing person detection error. Even if increase tendency on Re-Init rate is observed, overall change is less than 0.04% during experiment. This shows us power of CIGT motion modeling.


**Figure 27: Person detection effect on individual tracking by means of Re-Init metric.**

In last evaluation of person detection effect on individual tracking, we evaluate the positional errors by measuring RMSE [px] with respect to different error rate shown in Figure 28. RSME metric starts to increase after error rate is 0.10. However, overall increase is less than 1.5 pixels.


**Figure 28: Person detection effect on individual tracking by means of RSME (px) metric.**

### 5.4.2 Person Detection Effect on Group Detection and Tracking

Group detection is evaluated with 1-FP, 1-FN and GDSR metrics while group tracking is evaluated by means of MOTP (px) and MOTA.

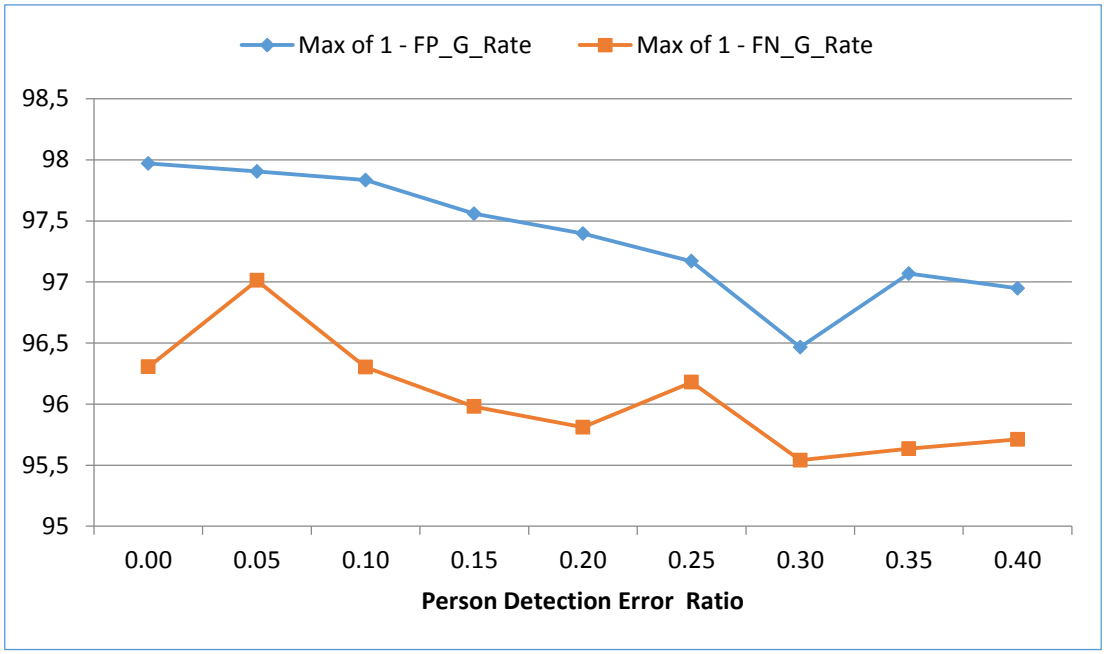



**Figure 29: Person detection effect on 1-FP and 1-FN metrics for group detection**

As shown in Figure 29, there is no significant change on 1-FP and 1-FN rates with respect to increasing error rate because of multi-observation model in CIGT. The false positive elimination mechanism in CIGT reduces the number of false positive for individuals and groups. As a result of this, 1-FP metric is less affected compared to 1-FN.

The other group detection metric is the GDSR that we evaluate the person detection effect on. According to Figure 30, GDSR does not significantly change. Since person detection simulator generates detection responses with random process, it is normal to observe small increase on decrease on GDSR result. However, overall change in GDSR metric is about 2%. As a result, CIGT framework is robust to increase in error rate for group detection.


**Figure 30: Person detection effect on GDSR for group detection**

For group tracking, we evaluate MOTP [m] and MOTA changes with respect to increasing error rate shown in Figure 31 and Figure 32, respectively. Since 1-FN rate in individual tracking decreases, some individual in group is identified, center position of group is drifted. As a result of this, MOTP increases and MOTA decreases while error rate increases. However, both MOTP [m] and MOTA are higher compared to similar trackers'.

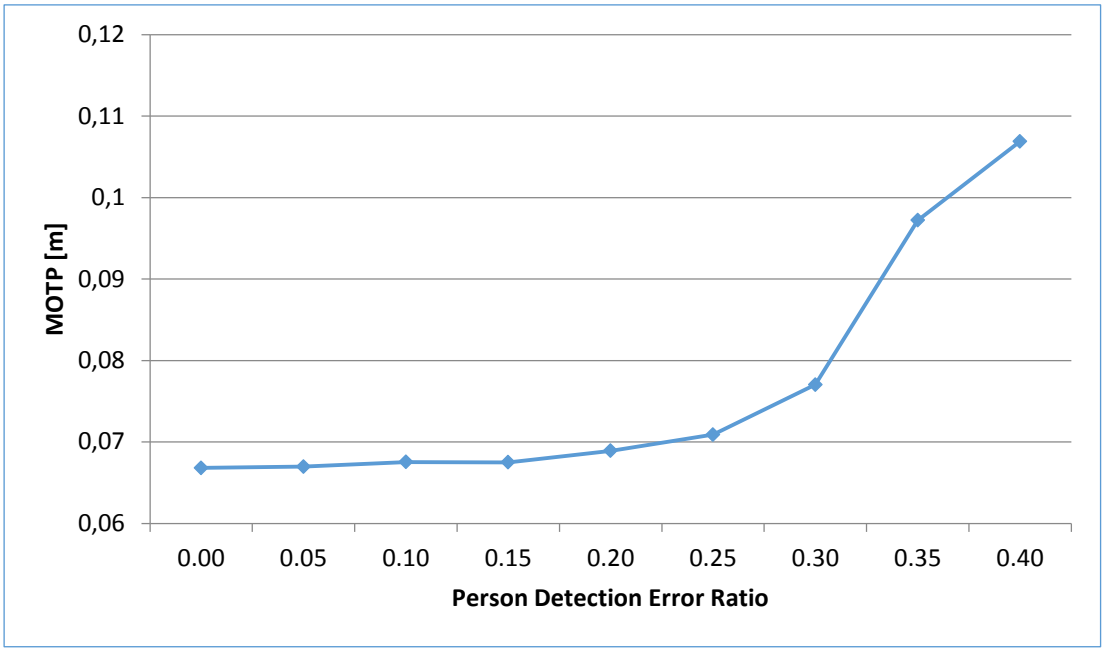



**Figure 31: Person detection effect on MOTP [m] for group tracking**

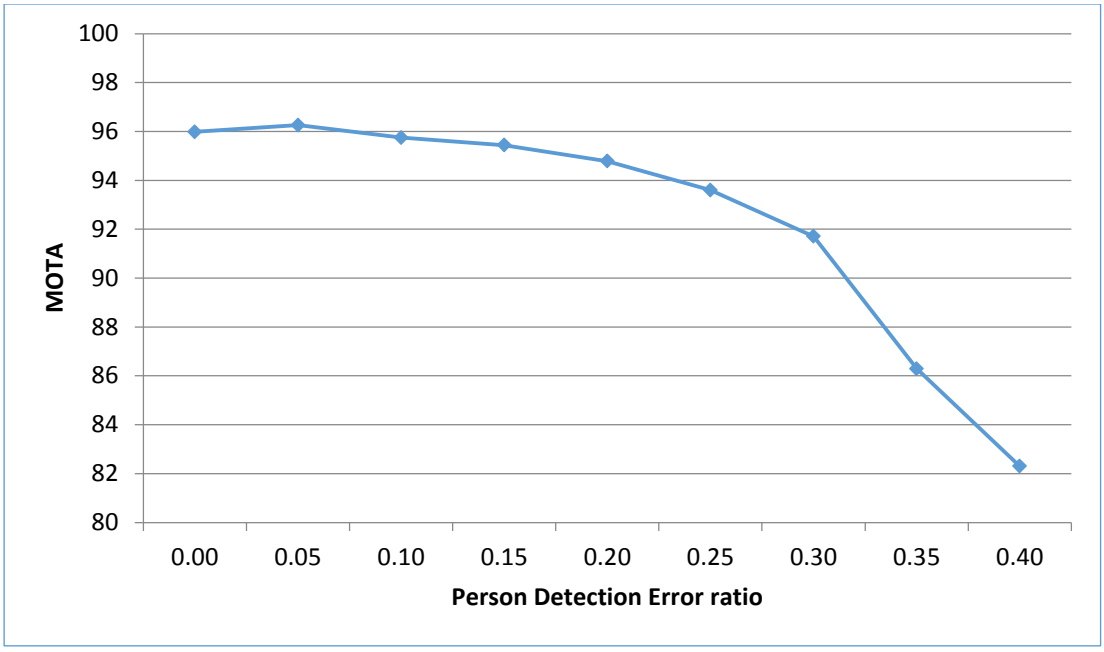



**Figure 32: Person detection effect on MOTA for group tracking**

In summary, motion model, two-phase association and false positive elimination mechanism in CIGT framework improve individual tracking performance. Also, multi-observation model forms groups and evaluates grouping events successfully and consequently it improves group detection and tracking accuracy with increasing person detection error rate.

## 5.5 Dynamic Motion Weight Evaluation

In this section, we provide the evaluation of dynamic motion weight parameter proposed in CIGT framework. In order to analyze the effects of dynamic parameters, we use video S2L1-1 in PETS 2009 dataset [77]. In this video, group density increases with respect to time shown in Figure 33 and dynamic parameter adjusts motion weights between template detector and particle advection.

**Figure 33: Increasing group density scenario on PETS 2009 [77]**

In this experiment, we conduct tests with edge values of dynamic parameter. Therefore, we evaluate dynamic parameter by comparing with 0 and 1 values. 0 means that only particle advection is source of motion model while 1 means that only template detector is used in motion model. The particle advection is more effective to model the motion pattern of dense groups while template matching is more feasible in sparse groups. Our proposed dynamic model adjusts the relative weights of these components and it is selected to be inversely proportional to group density. Figure 34 shows the effects of dynamic parameter on MOTP [px].

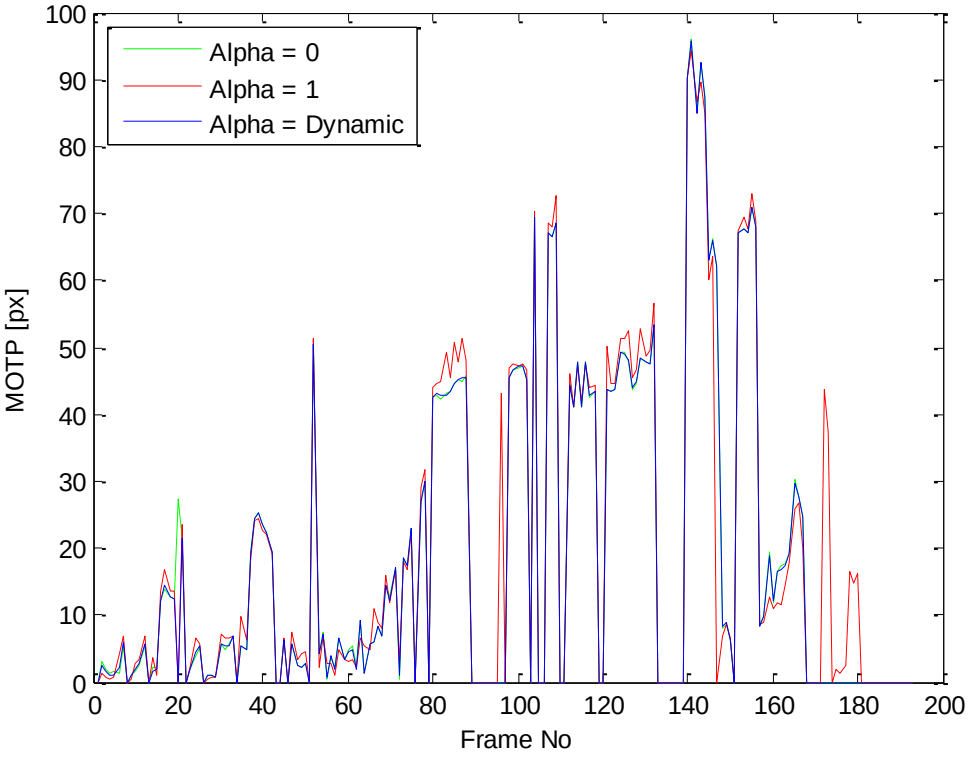



**Figure 34: Dynamic parameter evaluation**

At the beginning of video, group density is low and template detector can be more reliable than particle advection. Especially, around frame 20, particle advection has higher positional error compared to template detector. In this case, our proposed motion model automatically adjusts the motion weight close to template detector. However, in further frame (especially 80-90 and after 160), accuracy of particle advection increases since group gets denser. In this case, our motion model increases the weight of particle advection. However, since our modified template detector gives us good performance, there is no significant improvement observed with dynamic model.

## 5.6 Performance Evaluation

In this section, we evaluate the performance of CIGT framework by means of CPU usage, memory utilization and timing. In this experiment, we use Intel(R) Core™ i7-4770 CPU @ 3.5 GHz, 32 GB RAM and 64-bit Windows 7 PC and perform test on real part of FM dataset [7, 43] excluding queue scenarios. Table 19 shows the common processing performance results.

**Table 19: Processing performance metrics of CIGT framework on real part of FM dataset [7, 43] excluding queue scenarios.**

| Video | Average Number of Mid-level association per object | Average Frame Processing Time (seconds) | Total Memory Usage (MB) |
|---|---|---|---|
| **S01** | 3.37 | 12.83 | 2048.00 |
| **S02** | 12.24 | 41.54 | 1796.83 |
| **S03** | 0.98 | 7.74 | 1861.14 |
| **S04** | 3.31 | 11.68 | 1740.09 |
| **S05** | 11.44 | 37.92 | 1156.68 |
| **S06** | 6.75 | 20.64 | 1552.68 |
| **S07** | 10.71 | 35.98 | 1567.65 |
| **S08** | 2.04 | 9.93 | 1023.39 |
| **S09** | 5.96 | 18.51 | 1240.14 |
| **S10** | 7.15 | 21.15 | 957.32 |
| **S11** | 3.33 | 12.06 | 1232.17 |
| **S14** | 10.65 | 35.34 | 2048.00 |
| **S15** | 16.56 | 80.17 | 2042.45 |
| **Average** | **7.27** | **26.58** | **1558.96** |

We implement CIGT framework as single threaded application and we process each frame sequentially. When number of people increases in scenario, CIGT frame processing time increases gradually since all interactions between people are evaluated by CIGT framework. Due to this reason, we limit the maximum group size as 40. Number of interactions can be formulated as follows:

$$Number\ of\ Interaction = \binom{n}{2} \tag{5.12}$$

where $n$ is the number of people in scenario. Also, number of grouping events (merge and split) and mid-level association affects the frame processing time. Figure 35 shows how mid-level association count affects the CIGT frame processing time.

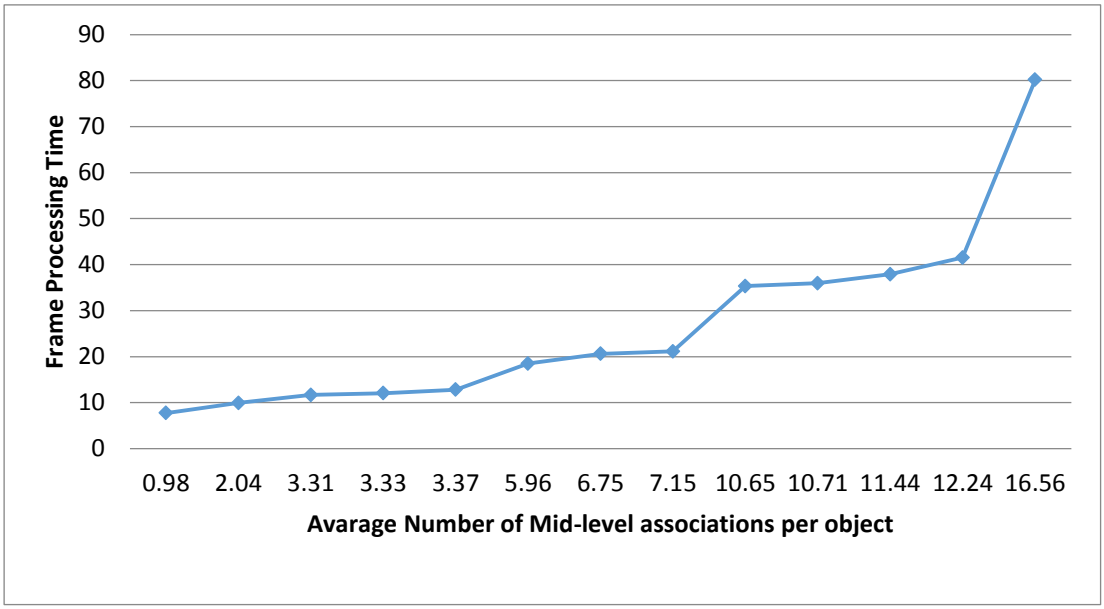



**Figure 35: Effect of mid-level association count on frame processing time in CIGT framework**

According to Table 19, memory utilization is over 1 GB on the average. In CIGT framework, we keep all frames for region covariance calculation and tracked object data (detected area, velocity, particles, etc.…) in order to perform association between objects. The advantage of this approach is to find object even if it is lost for some frames. However, keeping these data increases the memory utilization.

Our implementation is a single threaded application running on single core. Although our test PC has four physical cores in its architecture, eight cores are observed by using hyper threading. Therefore, CPU usage changes between 12.61%-12.82%.

In DPF based approaches [7, 43], object state is decomposed into individual state and group state. For each individual, certain number of particles are distributed randomly to estimate individual state. Also, certain number of particles are distributed for each particle in individual state estimate. As a result, number of particles is computed at each frame as follows:

$$p_{total} = p.n_i.n_g \tag{5.13}$$

where $p$ is the number of individuals, $n_i$ is the number of particles to estimate individual state, $n_g$ is the number of particles to estimate group state and $p_{total}$is the total number of particles in each frame.

Since the number of particles in DPF based approaches [7, 43] is higher than CIGT framework, computational cost of DPF based approaches [7, 43] is expected to be higher compared to CIGT framework. Since there is no association mechanism in [7, 43], the execution time of CIGT is probably higher than [7, 43] due to the cost of two-phase association. However, association increases the tracker accuracy in spite that it increases time complexity.

# CHAPTER 6

# CONCLUSION AND FUTURE WORKS

In this thesis, we combine sociological definitions with particle filtering framework and develop a new individual and group tracking methodology. Unlike standard particle filtering, our CIGT framework provides the flexibility to evaluate multiple observation measures. This observation modeling provides the ability of evaluating social interactions between people, analyzes the group events (merge and split) in group formation. In the observation model of CIGT, we evaluate not only distance but also direction of individuals. By this way, group crossing and groups approaching each other are also automatically evaluated in group formation and this feature enhances the group detection success. Different from DEEPER-JIGT factorization [7, 43], our CIGT framework does not decompose state of tracking objects into sub-states and keeps equivalence state information for groups and individuals. This factorization method reduces the system complexity since particles are sampled for only one state instead of multiple states and interaction between different states is omitted. Our CIGT framework uses particle advection to model motion pattern of tracking object. Instead of sparse optical flow, particle advection uses dense optical flow. Dense optical flow provides higher number of candidate particles to model tracking object's motion pattern. Our CIGT framework uses hierarchical methods to find correct particles for motion modeling. These mechanisms provide better positioning the tracking object and as a result, spatial error for group tracking is lower than similar tracking methods in state-of-the art.

In addition, our CIGT framework combines online learning model based on discriminative appearance model [5] with the tracking engine. Discriminative appearance model defines object with not only color but also texture and shape features. This provides us to associate objects with correct match and increase the rate of correct state estimate of tracking object. False positive detections negatively affect the tracker performance. To alleviate its negative effect, we use False Positive elimination mechanism to reduce the number of false positive detections and tracking objects. This false elimination mechanism provides the better detection of groups and reduces the effects of detection errors in tracking.

A summary of pros and cons of the proposed approach are as follows:

- Two-phase association model in suggested method performs association between objects at different frames and increases the state estimate of tracking object and id switch count. Also, it helps our proposed detection correlation method in the false positive elimination mechanism and AdaBoost online trainer with spatial constraints.

- Our CIGT framework proposes a hierarchical approach to eliminate false positive and increase tracker performance, reduce id switch and wrong estimation of group state. However, bad illumination and noises may cause the poor appearance model of tracking objects or foreground objects may not be identified. As a result of these bad conditions' effects, this mechanism in CIGT may eliminate the true detections and increase the false negative rate.

- The power of suggested tracking framework comes from multi-observation model inspired from sociological definitions of in-groups and out-groups. This observation model provides us to evaluate not only appearance similarity but also interactions between people in group. As a result of this, grouping events are evaluated in tracking framework with multiple weights. The other advantage of multi-observation model is to be expandable. New interaction types can be added into multi-observation model and this observation model can be used in different works other than tracking individuals and groups.

- The CIGT framework proposes hierarchical approach to model tracking objects' motion pattern. This method provides us the better estimating velocity of tracking object and positioning the tracking objects.

Our proposed CIGT framework is tested in 46 videos which consist of 25 synthetic, 21 real sequences. In the experiments, we first observe that two-phase association shows great performance at reducing id switch and helping the group state estimate. ID switch is very important for individual tracking and it directly affects the group state estimate. Therefore, reducing id switch increases not only the individual tracking performance but also the group detection and tracking results. In our proposed framework, number of id switches is remarkably low compared to similar methods. We also observe that group detection success is high-

er than DP2-JIGT [43] even if our suggested method does not include online inference for group formation.

In addition, we evaluate the effect of person detection errors on our proposed framework. In this experiment, we first observe that false positive elimination mechanism greatly reduces the false positive detections without being adversely affected by the person detection error rate. Although similar methods have higher false positive rates, the proposed framework remains robust against increasing person detection rate.

A future direction of research is to focus on online inference methodology for group formation. Our literature survey indicates that CRF model is suitable to use in multi-observation model. In CRF model, setting the unary terms as individuals' global appearance similarity and pairwise term as social interactions the energy function for group formation can be built. Then, our problem turns into an energy minimization function.

Recall that multi-observation model has expandable property. In the future, we have a plan to use model in crowd analysis. We add new observation from different interactions so that abnormal events can be detected during tracking process. Besides multi-observation model, we will analyze the flows of groups or crowds during tracking process since our model supports particle advection.

Our implementation is single threaded and runs on a single CPU core. In the future, it can be implemented on GPU for real-time applications. In the current implementation, particles are evaluated sequentially. However, particles in each frame can be evaluated independently and we can run each particle weight calculation in parallel on GPU. In addition, region covariance is calculated for each object sequentially and computed each time. However, we can compute region covariance separately on GPU and keep in the memory to reduce the computation time.

# REFERENCES


[1] J. Jacques Junior, S. Raupp Musse and C. Jung, "Crowd Analysis Using Computer Vision Techniques," *Signal Processing Magazine, IEEE*, pp. 66-77, 2010.

[2] A. Yigit and A. Temizel, "Particle filter based Conjoint Individual-Group Tracker (CIGT)," in *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*, 2015.

[3] S. Ali and S. Mubarak, "A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1-6, 2007.

[4] B. Solmaz, B. E. Moore and M. Shah, "Identifying behaviors in crowd scenes using stability analysis for dynamical systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 2064-2070, 2012.

[5] C. H. Kuo, C. Huang and R. Nevetia, "Multi-target tracking by on-line learned discriminative appearance models," *Cvpr*, pp. 685-692, 2010.

[6] H. Arrow, J. E. McGrath and J. L. Berdahl, Small groups as complex systems: Formation, coordination, development, and adaptation, Sage Publications, 2000.

[7] L. Bazzani, M. Cristani and V. Murino, "Decentralized particle filter for joint individual-group tracking," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1886-1893, 2012.

[8] M. Rodriguez, I. Laptev, J. Sivic and J.-Y. Audibert, "Density-aware Person Detection and Tracking in Crowds," *Proceedings of the 2011 International Conference on*

*Computer Vision*, pp. 2423--2430, 2011.

[9] I. Ali and M. N. Dailey, "Multiple human tracking in high-density crowds," *Advanced Concepts for Intelligent Vision Systems*, pp. 540-549, 2009.

[10] C. Beyan and A. Temizel, "Adaptive mean-shift for automated multi object tracking," *Computer Vision, IET*, vol. 6, pp. 1-12, 2012.

[11] C. Huang, B. Wu and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," *Eccv*, vol. 5303 LNCS, no. PART 2, pp. 788-801, 2008.

[12] B. Yang and R. Nevatia, "Online learned discriminative part-based appearance models for multi-human tracking," in *Computer Vision-ECCV 2012*, 2012.

[13] S.-H. Bae and K.-J. Yoon, "Robust Online Multiobject Tracking With Data Association and Track Management," *Image Processing, IEEE Transactions on*, vol. 23, no. 7, pp. 2820-2833, 2014.

[14] A. Andriyenko and K. Schindler, "Multi-target tracking by continuous energy minimization," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011.

[15] A. Andriyenko, K. Schindler and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012.

[16] A. Milan, S. K. and S. Roth, "Detection- and Trajectory-Level Exclusion in Multiple Object Tracking," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013.

[17] B. Yang and R. Nevatia, "An Online Learned CRF Model for Multi-Target Tracking," *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2034-2041, 2012.

[18] B. Yang and R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012.

[19] M. Luber, J. Stork, G. Tipaldi and K. Arras, "People tracking with human motion predictions from social forces," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, 2010.

[20] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier and L. Van Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009.

[21] L. Sun, G. Liu and Y. Liu, "Multiple pedestrians tracking algorithm by incorporating histogram of oriented gradient detections," *Image Processing, IET,* vol. 7, pp. 653-659, 2013.

[22] X. Wang and Z. Tang, "Modified particle filter-based infrared pedestrian tracking," *Infrared Physics & Technology,* vol. 53, no. 4, pp. 280-287, 2010.

[23] L. Kratz and K. Nishino, "Tracking Pedestrians Using Local Spatio-Temporal Motion Patterns in Extremely Crowded Scenes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 34, no. 5, pp. 987-1002, 2012.

[24] Y. He, M. Pei, M. Yang, Y. Wu and Y. Jia, "Online visual tracking by integrating spatio-temporal cues," *Computer Vision, IET,* vol. 9, no. 1, pp. 124-137, 2015.

[25] S. Mittal, T. Prasad, S. Saurabh, X. Fan and H. Shin, "Pedestrian detection and tracking using deformable part models and Kalman filtering," in *SoC Design Conference (ISOCC), 2012 International*, 2012.

[26] P. Felzenszwalb, D. McAllester and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008.

[27] H. Zhou, Y. Yuan and C. Shi, "Object tracking using SIFT features and mean shift," *Computer Vision and Image Understanding,* vol. 113, no. 3, pp. 345-352, 2009.

[28] M. Andersson and J. Rydell, "Crowd analysis with target tracking, K-means clustering and hidden Markov models," *15th Int'l Conf. on Information Fusion (FUSION),* pp. 1903-1910, 2012.

[29] L. Bazzani, M. Cristani and V. Murino, "Collaborative particle filters for group tracking," *Image Processing (ICIP), 2010 17th IEEE International Conference on,* pp. 837-840, 2010.

[30] A. Setia and A. Mittal, "Co-operative Pedestrians Group Tracking in Crowded Scenes Using an MST Approach," *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on,* pp. 102-108, 2015.

[31] R. Mazzon, F. Poiesi and A. Cavallaro, "Detection and tracking of groups in crowd," *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on,* pp. 202-207, 2013.

[32] F. Poiesi and A. Cavallaro, "Detection and tracking of interacting targets," in *IEEE Trans. on IP, (submitted),* 2013.

[33] T. Linder and K. Arras, "Multi-model hypothesis tracking of groups of people in RGB-D data," *Information Fusion (FUSION), 2014 17th International Conference on,* pp. 1-7, 2014.

[34] S. Zaidenberg, B. Boulay, C. Garate, D. P. Chau, E. Corvee and F. Bremond, "Group interaction and group tracking for video-surveillance in underground railway stations," *International Workshop on Behaviour Analysis and Video Understanding (ICVS 2011),* 2011.

[35] T. Ojala, M. Pietikainen and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, 1994.

[36] C. Garate, S. Zaidenberg, J. Badie and F. Bremond, "Group tracking and behavior recognition in long video surveillance sequences," *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on,* vol. 2, pp. 396-402, 2014.

[37] S. K. Pang, J. Li and S. Godsill, "Models and Algorithms for Detection and Tracking of Coordinated Groups," in *Aerospace Conference, 2008 IEEE*, 2008.

[38] G. Gennari and G. Hager, "Probabilistic data association methods in visual tracking of groups," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2004.

[39] B. Lau, K. Arras and W. Burgard, "Tracking groups of people with a multi-model hypothesis tracker," in *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, 2009.

[40] P. Shea, K. Alexander and J. Peterson, "Group tracking using genetic algorithms," in *Information Fusion, 2003. Proceedings of the Sixth International Conference of*, 2003.

[41] V. Edman, M. Andersson, K. Granstrom and F. Gustafsson, "Pedestrian group tracking using the GM-PHD filter," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, 2013.

[42] D. Ryan, S. Denman, C. Fookes and S. Sridharan, "Crowd Counting Using Group Tracking and Local Features," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, 2010.

[43] L. Bazzani, M. Zanotto, M. Cristani and V. Murino, "Joint Individual-Group Modeling for Tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 4, pp. 746-759, 2015.

[44] T. Chen, T. Schon, H. Ohlsson and L. Ljung, "Decentralized Particle Filter With Arbitrary State Decomposition," *Signal Processing, IEEE Transactions on*, vol. 59, no. 2, pp. 465-478, 2011.

[45] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vols. PAMI-1, no. 2, pp. 224-227, 1979.

[46] X. Gu, J. Cui and Q. Zhu, "Abnormal crowd behavior detection by using the particle entropy," *Optik - International Journal for Light and Electron Optics*, vol. 125, no. 14, pp. 3428-3433, 2014.

[47] R. Mehran, A. Oyama and M. Shah, "Abnormal crowd behavior detection using social force model," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE*

*Conference on*, 2009.

[48] B. Krausz and C. Bauckhage, "Automatic detection of dangerous motion behavior in human crowds," in *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, 2011.

[49] H. Ullah and N. Conci, "Crowd motion segmentation and anomaly detection via multi-label optimization," in *ICPR workshop on Pattern Recognition and Crowd Analysis*, 2012.

[50] R. Raghavendra, A. Del Bue, M. Cristani and V. Murino, "Optimizing interaction force for global anomaly detection in crowded scenes," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 2011.

[51] C. Ongun, A. Temizel and T. Temizel, "Local anomaly detection in crowded scenes using Finite-Time Lyapunov Exponent based clustering," in *Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on*, 2014.

[52] S. Ali and M. Shah, "Floor Fields for Tracking in High Density Crowd Scenes," in *In Proc. of European Conf on Computer Vision*, 2008.

[53] H. Idrees, N. Warner and M. Shah, "Tracking in Dense Crowds Using Prominence and Neighborhood Motion Concurrence," *Image Vision Comput.*, vol. 32, no. 1, pp. 14-26, 2014.

[54] A. Dehghan, H. Idrees, A. R. Zamir and M. Shah, "Automatic Detection and Tracking of Pedestrians in Videos with Various Crowd Densities," in *Pedestrian and Evacuation Dynamics 2012*, Springer, 2014, pp. 3-19.

[55] D. R. D. Forsyth, Group Dynamics, 2009.

[56] D. Helbing and P. Molnar, "Social Force Model for Pedestrian Dynamics," *Physical Review E,* vol. 51, no. 5, pp. 4282-4286, 1998.

[57] R. Privman, S. R. Hiltz and Y. Wang, "In-Group (Us) versus out-group (Them) dynamics and effectiveness in partially distributed teams," *IEEE Transactions on Professional Communication*, vol. 56, no. 1, pp. 33-49, 2013.

[58] S. Pellegrini, K. Schindler and L. van Gool, "You'll Never Walk Alone: Modeling Social Behavior for Mluti-target Tracking," *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV)*, no. Iccv, pp. 261-268, 2009.

[59] K. N. Tran, A. Bedagkar-Gala, I. A. Kakadiaris and S. K. Shah, "Social Cues in Group Formation and Local Interactions for Collective Activity Analysis," *VISAPP (1)*, pp. 539-548, 2013.

[60] R. E. Schapire and R. E. Schapire, "Improved Boosting Algorithms Using Con dence-rated Predictions," *Machine Learning*, vol. 37, no. 3, pp. 297-336, 1999.

[61] O. Tuzel, F. M. Porikli and P. Meer, "Region Covariance: A Fast Descriptor for Detection and Classification," *Europe Conf. Comp. Vision (ECCV)*, vol. II, pp. 589-600, 2006.

[62] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pp. 886-893, 2005.

[63] M. H. Bharati, J. Liu and J. F. MacGregor, "Image texture analysis: methods and comparisons," *Chemometrics and Intelligent Laboratory Systems*, vol. 72, no. 1, pp. 57-71, 2004.

[64] R. Haralick, K. Shanmugam and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans. on Systems, Man and Cybernetics*, pp. 610-621, 1973.

[65] M. M. Galloway, "Texture analysis using gray level run lengths," *Comput. Graphics Image Process*, vol. 4, pp. 172-179, 1975.

[66] L. Carlucci, "A formal system for texture languages," *Pattern Recognition*, vol. 4, pp. 53-72, 1972.

[67] A. Sarkar, K. Sharma and R. Sonak, "A new approach for subset 2-D AR model identification for describing textures," *Image Processing, IEEE Transactions on*, vol. 6, no. 3, pp. 407-413, 1997.

[68] J. Ning, L. Zhang, D. Zhang and C. Wu, "Robust Object Tracking Using Joint Color-

Texture Histogram," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 23, pp. 1245-1263, 2009.

[69] P. Geladi, "Some special topics in multivariate image analysis," *Chemometrics and Intelligent Laboratory Systems,* vol. 14, no. 1, pp. 375-390, 1992.

[70] A. Bovik, M. Clark and W. Geisler, "Multichannel texture analysis using localized spatial filters," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 12, no. 1, pp. 55-73, 1990.

[71] B. Drayer and T. Brox, "Training Deformable Object Models for Human Detection Based on Alignment and Clustering," in *Computer Vision – ECCV 2014*, 2014.

[72] L. Spinello and K. Arras, "People detection in RGB-D data," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, 2011.

[73] S. Stalder, H. Grabner and L. Van Gool, "Cascaded Confidence Filtering for Improved Tracking-by-Detection," in *Computer Vision – ECCV 2010*, 2010.

[74] J. Yao and J.-M. Odobez, "Multi-Layer Background Subtraction Based on Color and Texture," *2007 IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1-8, 2007.

[75] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *IJCAI,* vol. 81, pp. 674-679, 1981.

[76] Z. Kalal, K. Mikolajczyk and J. Matas, "Forward-backward error: Automatic detection of tracking failures," *Proceedings - International Conference on Pattern Recognition,* pp. 2756-2759, 2010.

[77] PETS2009, "Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance.," 2009.

[78] M. Andriluka, S. Roth and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008.

[79] K. Smith, D. Gatica-Perez, J. Odobez and S. Ba, "Evaluating Multi-Object Tracking," in *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, 2005.

[80] A. Milan, K. Schindler and S. Roth, "Challenges of Ground Truth Evaluation of Multi-target Tracking," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, 2013.

[81] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *J. Image Video Process.*, vol. 2008, pp. 1-10, 2008.

# CURRICULUM VITAE

## PERSONAL INFORMATION

**Surname, Name:** Yiğit, Ahmet
**Nationality:** Turkish (TC)
**Date and Place of Birth:** 2 March 1982, Adana
**Email:** ahmetyigit@gmail.com

## EDUCATION

| Degree | Institution | Year of Graduation |
|---|---|---|
| MS | METU, Department of Information Systems | 2010 |
| BS | METU, Department of Computer Engineering | 2005 |

## WORK EXPERIENCE

| Year | Organization | Position |
|---|---|---|
| 2005- Present | HAVELSAN AŞ., Ankara | Senior Software Engineer |

## PUBLICATIONS

[1] A. Yigit and A. Temizel, "Individual and Group Tracking with the Evaluation of Social Interaction", IET Computer Vision, (Under preparation)

[2] A. Yigit and A. Temizel, "Particle filter based Conjoint Individual-Group Tracker (CIGT)," in Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on, 2015.

[3] C.Beyan, A.Yigit, A.Temizel, "Fusion of Thermal and Visible Band Video for Abandoned Object Detection", *J. Electron. Imaging*. 20(3), 033001 (July 08, 2011)

[4] A.Yigit, A.Temizel, "Termal ve Görünür Bant İmgeleri Kullanılarak Terk Edilen Nesne Tespiti", IEEE Signal Processing, Communication and Applications Conference, April 2010.

**TEZ FOTOKOPİSİ İZİN FORMU**

**ENSTİTÜ**

| | |
|---|---|
| Fen Bilimleri Enstitüsü | |
| Sosyal Bilimler Enstitüsü | |
| Uygulamalı Matematik Enstitüsü | |
| Enformatik Enstitüsü | X |
| Deniz Bilimleri Enstitüsü | |

**YAZARIN**

Soyadı : YİĞİT
Adı : AHMET
Bölümü : BİLİŞİM SİSTEMLERİ

**TEZİN ADI** (İngilizce): CONJOINT INDIVIDUAL AND GROUP TRACKING FRAMEWORK WITH ONLINE LEARNING

**TEZİN TÜRÜ** : Yüksek Lisans [ ] Doktora [X]

1. Tezimin tamamından kaynak gösterilmek şartıyla fotokopi alınabilir. [X]
2. Tezimin içindekiler sayfası, özet, indeks sayfalarından ve/veya bir bölümünden kaynak gösterilmek şartıyla fotokopi alınabilir. [ ]
3. Tezimden bir (1) yıl süreyle fotokopi alınamaz. [ ]

**TEZİN KÜTÜPHANEYE TESLİM TARİHİ**: