IET Computer Vision

Research Article

IET Journals
The Institution of Engineering and Technology

Individual and group tracking with the evaluation of social interactions

ISSN 1751-9632
Received on 15th July 2016
Revised 5th December 2016
Accepted on 6th December 2016
doi: 10.1049/iet-cvi.2016.0238
www.ietdl.org

Ahmet Yigit[1] ✉, *Alptekin Temizel*[1]

[1]*Graduate School of Informatics, Middle East Technical University, Ankara, Turkey*
✉ *E-mail: e125095@metu.edu.tr*

Abstract: Tracking groups of people is a challenging problem. Groups may grow or shrink dynamically with merging and splitting of individuals and conventional trackers are not designed to handle such cases. In this study, the authors present a conjoint individual and group tracking (CIGT) framework based on particle filter and online learning. CIGT has four complementary phases: two-phase association, false positive elimination, tracking and learning. First, reliable tracklets are created and detection responses are associated to tracklets in two-phase association. Then, hierarchal false positive elimination is performed for unassociated detection responses. In the tracking phase, CIGT calculates multiple weights from the observation and jointly models individuals and groups. Particle advection is used in the motion model of CIGT to facilitate tracking of dense groups. In the learning phase, the discriminative appearance model, consisting of shape, colour and texture features, is extracted and used in AdaBoost online learning. Using the discriminative learning model, state estimation is performed on both individuals and groups. The experimental results show that the performance of the proposed framework compares favourably with other individual and group-tracking methods for both real and synthetic datasets.

1 Introduction

People-tracking plays an important role in video surveillance. Although many challenges have been addressed, there are still open challenges in dense environments and when there are groups of people.

A group is a social unit; persons in a group interact with each other and share similar characteristics. Groups are dynamic units which may grow or shrink as a result of merging or splitting of individuals. Due to these challenges, tracking of groups requires consideration of not only visual but also social properties such as the interaction between the individuals.

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

2 Related work and background

Although individual-tracking and group-tracking methods address different problems, individuals and groups co-exist in various scenes, necessitating both group and individual tracking methods to be used together. Recently, there has been a growing interest in handling individual and group-tracking problems in a single framework. In this section, we first summarise the related work and highlight the differences of the proposed method in comparison with the state-of-the-art. Then we provide background information on the sociological aspects used in formulating the proposed method.

2.1 Individual and group-tracking methods

Sparse and dense crowds could be analysed by using separate approaches. In the microscopic approach, individuals are evaluated separately and multi target tracking is used. On the other hand, dense crowds are tracked using the macroscopic approach where a group tracking method is used [1]. Decentralised particle filter [2] based approaches decompose the object state into two sub-states: a group label, to which the individual belongs, and individual velocity and position [3, 4]. Then the posterior distribution is factorised as follows:

$$p(Z_t, X_{0:t}|y_{0:t}) = p(Z_t|X_{0:t}, y_{0:t})p(X_{0:t}|y_{0:t}) \quad (1)$$

where Z_t is the group state estimate, $X_{0:t} = (X_0, X_1, \ldots, X_t)$ is the set of individual state estimate, and $y_{0:t} = (y_0, y_1, \ldots, y_t)$ is the set of observations. Individual state estimates $X_{0:t}$ are fed to group state estimate Z_t. An online inference mechanism for group formation based on the Dirichlet process mixture models eliminates the need to explicitly model the group events such as merge and split [4]. However, lack of association between consecutive frames and online learning mechanism for individual tracking consequently results in high number of ID switches [3, 4]. The state model separates the individual state from group state [3, 4] and this increases the complexity of the system. The group state estimate uses only the weighted colour histogram and other features such as texture and shape are not considered.

Particle advection method can be used to analyse the motion flows of dense crowds and perform stability analysis [5, 6]. It has been shown to be effective in analysing crowd dynamics especially when the crowd density is high and the tracking of individuals is not feasible due to occlusions. Recently, a number of methods have been proposed to make use of particle advection in tracking [7, 8]. These methods use particle advection to analyse the crowd motion patterns and track in dense crowds. However, no explicit mechanism is provided to extract the motion patterns for individual and group tracking.

In this paper, we propose a tracking method based on a multi-observation model to handle group events such as merge and split. Different to the methods in [3, 4], we use two-phase association and online learning model for individual tracking that reduces the number of ID switches. Discriminative appearance model [9] is embedded into the state estimate and it is used to identify individuals and groups. Different to the methods in [7, 8], we propose a motion extraction scheme by using particle advection for individual and group tracking.

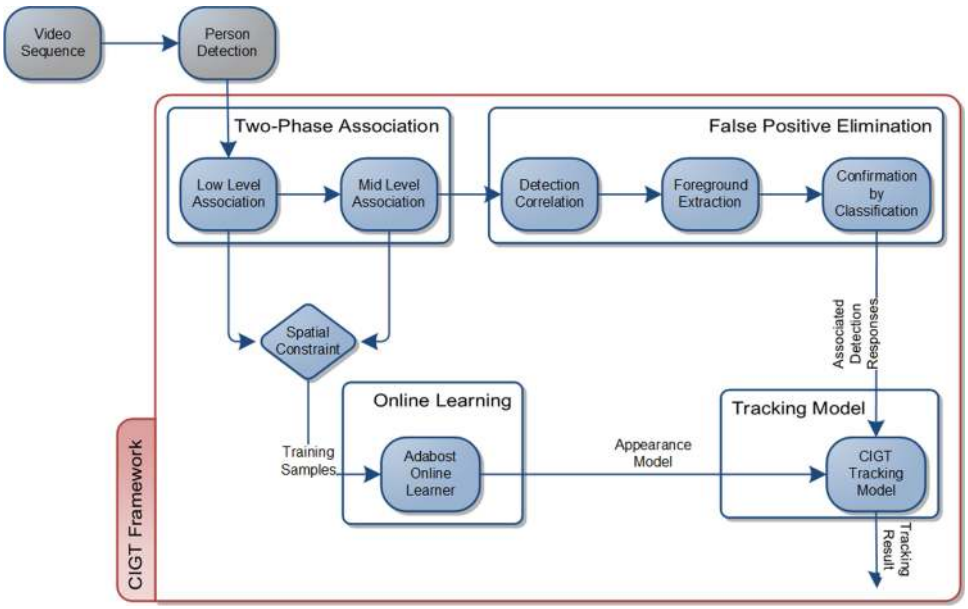

Fig. 1 *Proposed framework*

2.2 Sociological background

A group is defined as two or more individuals who are related to each other by social relationships. The basic characteristics of groups are interaction, goals, interdependence, structure and unity [10].

To achieve common goals, individuals in a group perform certain interactions but these interactions should be consistent and should not break the unity feature of groups. In addition, interaction affects group formation, structure, and interdependency between the group members. Since interaction is a measurable property, most methods in the literature use it in group analysis and tracking. For example, the interactions between pedestrians can be modelled in terms of repulsive and attractive forces called social forces [11]. Based on these two forces, an interaction energy function can be modelled for multi-target tracking [12].

Since a group is a dynamic entity, interaction not only between the group members but also between group members and other individuals should be taken into account in the process of group tracking. In order to evaluate these interactions, individuals in a group can be analysed in terms of in-group and out-group measurements. In sociology, an in-group is defined as a social group, with which a person identifies as being a member. By contrast, an out-group is a social group with which an individual does not identify. Inspired by these definitions, we introduce in-group and out-group measures to evaluate interactions.

Interactions can be analysed as a function of distance and angular displacement between people [12]. The unity feature of group means that the group members are close enough to each other and have similar angular motion directions. The possibility of interactions tends to increase when people get closer to each other [13]. We make use of these features of groups in formulation of our framework to evaluate social interactions.

3 Conjoint individual and group tracking framework

The proposed method is partly based on our earlier work [14] and differs from the existing methods in the literature in that it holds joint state information for both groups and individuals, and evaluates group formation using the incorporated multi-observation model. Unlike the standard particle filter, the multi-observation model can also evaluate the interactions between individuals based on the defined in-group and out-group weights. Furthermore, to account for changing group size and density, the model dynamically controls the number of particles. In the motion model, both particle advection and template detection are used to track dense and sparse groups respectively. Furthermore, in the proposed method, a discriminative appearance model [9] and online learning are used to increase the performance and state estimation of groups.

The proposed framework consists of four main parts: two-phase association, online learning, false positive elimination and tracking model (Fig. 1). The inputs of the framework are raw video sequences and person detection results (bounding boxes of the detected person areas). The *two-phase association* model is used to associate individuals from previous frames (Section 3.1). A spatial constraint is performed on association results to create positive and negative training samples which are fed to the *online learning part* [15] (Section 3.2). *False positive elimination* part is used to reduce the false positive detections from the responses (Section 3.3). The *tracking model* incorporates a multi-observation model, a motion model and particle resampling phase (Section 3.4).

3.1 Two-phase association

This part aims to associate the independent person detection outputs in consecutive frames and identifies the ones belonging to the same object.

Low-level association: First, the affinity score matrix $\boldsymbol{S}$, elements of which consists of affinity scores calculated by multiplying position, size and colour histogram similarity measurements between detection responses r_i and r_j in consecutive frames is formed [16]. Then a dual-threshold method is applied and r_i and r_j are considered to belong to the same tracklet if the following conditions are satisfied [16]:

$$\begin{aligned} \boldsymbol{S}(i,j) &> \theta_1 \\ \forall x \in R - \{i\}, \boldsymbol{S}(i,j) - \boldsymbol{S}(x,j) &> \theta_2 \\ \forall y \in R - \{j\}, \boldsymbol{S}(i,j) - \boldsymbol{S}(i,y) &> \theta_2 \end{aligned} \tag{2}$$

where $R = \{r_i\}$ is the set of all detection responses, θ_1 and θ_2 are thresholds, $\boldsymbol{S}(i,j)$ is the affinity score between r_i and r_j. The dual-threshold strategy provides a conservative and biased way of linking only the reliable associations. Detection responses associated by low-level association are not passed onto the mid-level association. This reduces the computation time in mid-level association and increases its accuracy by reducing the search space. On the other hand, low-level association does not resolve the ambiguity of conflicting pairs and it is short-term (i.e. using only the consecutive frames).

Mid-level association: Mid-level association is a long-term association method and performed only for detections that cannot be associated in the low-level. Unlike low-level association, not

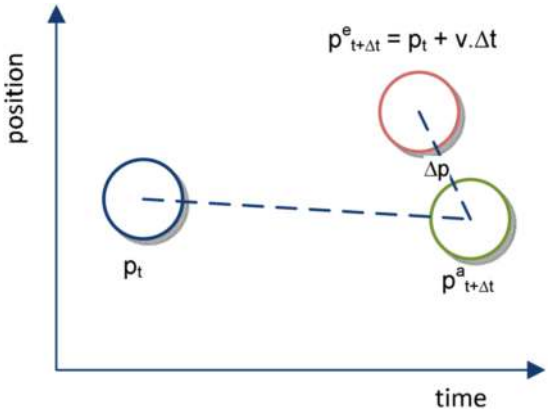

Fig. 2 *Linear motion model to estimate position*

only appearance but also spatial information is considered by separately calculating the appearance and the motion similarity scores.

We use the discriminative appearance model [9] which uses colour histogram, histogram of oriented gradients (HOG) [17] and region covariance [18] based features. Colour histograms are calculated by using 8 bins for each RGB channel. Then, these three vectors are concatenated to form a single 24-element vector $\boldsymbol{f}_{\mathrm{RGB}_i}$. A HOG feature vector $\boldsymbol{f}_{\mathrm{HOG}_i}$ is formed by concatenating eight orientation bins in 2×2 cells over the rectangle region R. The texture descriptor corresponding to the covariance matrix over R is defined as follows [18]:

$$\boldsymbol{C}_R = \frac{1}{n-1}\sum_{k=1}^{n}(\boldsymbol{z_k}-\mu)(\boldsymbol{z_k}-\mu)^{\mathrm{T}} \tag{3}$$

$$\boldsymbol{z}_k = \left[\frac{\partial I}{\partial x}\frac{\partial I}{\partial y}\frac{\partial^2 I}{\partial x^2}\frac{\partial^2 I}{\partial y^2}\frac{\partial^2 I}{\partial x \partial y}\right]^{\mathrm{T}} \tag{4}$$

where z_k consists of first and second order derivatives of intensity image I at kth pixel in R, μ is the mean over R, and n is the number of pixels in R.

Once the appearance model is formed, we can calculate the similarity measures between the non-associated detection responses and the tracklets found previously. We employ correlation coefficient for both colour and HOG features. The similarity between covariance matrices is measured as follows [19]:

$$\sigma(\boldsymbol{C}_i, \boldsymbol{C}_j) = \sqrt{\sum_{k=1}^{5}\ln^2(\lambda_k(\boldsymbol{C}_i, \boldsymbol{C}_j))} \tag{5}$$

where $\lambda_k(\boldsymbol{C}_i, \boldsymbol{C}_j)$ are the generalised eigenvalues of $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$:

$$\lambda_k \boldsymbol{C}_i \boldsymbol{x}_k - \boldsymbol{C}_j \boldsymbol{x}_k = 0, \quad k = 1, \ldots, 5 \text{ and } x_k \neq 0. \tag{6}$$

Then, the appearance similarity score is calculated:

$$\begin{aligned} P_{\mathrm{hist}} &= G\left(-1 + cc\left(\boldsymbol{f}_{\mathrm{RGB}_i}, \boldsymbol{f}_{\mathrm{RGB}_j}\right)\right) \\ P_{\mathrm{HOG}} &= G\left(-1 + cc\left(\boldsymbol{f}_{\mathrm{HOG}_i}, \boldsymbol{f}_{\mathrm{HOG}_j}\right)\right) \\ P_{RG} &= G\left(\sigma(\boldsymbol{C}_i, \boldsymbol{C}_j)\right) \\ S_a &= P_{RG} P_{\mathrm{hist}} P_{\mathrm{HOG}} \end{aligned} \tag{7}$$

where P_{hist}, P_{HOG} and P_{RG} are zero mean Gaussian function of distance values for the histogram, HOG and region covariance, respectively, G is the zero mean Gaussian, and cc is the normalised cross correlation coefficient functions.

As shown in Fig. 2, the motion similarity score is based on the distance between the position estimated with the linear motion model $p^e_{t+\Delta t} = p_t + v_t \cdot \Delta t$ and the real position $p^a_{t+\Delta t}$, where v_t is the velocity, p_t is the position at frame t, and Δt is the frame difference [20]. The position difference is calculated as $\Delta p = p^a_{t+\Delta t} - p^e_{t+\Delta t}$ and we compute the motion similarity score as follows:

$$S_m = G(\Delta p, \textstyle\sum_p) \tag{8}$$

where G is the zero mean Gaussian function with $\sum_p$ variance.

Finally, the probability of association which takes account of both appearance and motion similarities between objects is calculated as follows:

$$P_s = S_a \cdot S_m \tag{9}$$

The probability of association is calculated between objects and previously identified tracklets. Then, we select the tracklet having the highest probability of association as the candidate.

As a result of the two-phase association, associated object pairs are generated. These pairs are assumed to belong to the same object and passed onto the online learning part as positive training examples. Two phase association step is followed by utilisation of a spatial constraint for negative sample collection where we exploit the fact that an object cannot belong to two different tracklets at the same frame. Therefore, we create negative samples using pairs of associated objects which are spatially far away from each other. Moreover, we build a discriminative set for each tracklet [9] to collect negative samples.

3.2 False positive elimination

Conjoint individual and group tracking (CIGT) proposes a hierarchical method to reduce the number of false positives in detection results using three phases: detection correlation, foreground extraction and confirmation-by-classification [21]. Detection correlation aims to correlate detections not associated by two-phase association with other detections. Most of the detections are identified by two-phase association and our aim is to eliminate the detection of suddenly appearing objects, not satisfying the spatial constraint or intersecting with others. Therefore, we first compute the minimum intersection ratio between the two detections M_1 and M_2 where M' is their area of intersection:

$$I_s = \min\left(\frac{M'}{M_1}, \frac{M'}{M_2}\right) \tag{10}$$

Then, we compute the appearance similarity between them:

$$\begin{aligned} K_s &= P_{\mathrm{hist}} P_{\mathrm{HOG}} P_{\mathrm{size}} \\ P_{\mathrm{size}} &= G\left(\sqrt{\left(\frac{w_i - w_j}{\max\left(w_i, w_j\right)}\right)^2 + \left(\frac{h_i - h_j}{\max\left(h_i, h_j\right)}\right)^2}\right) \end{aligned} \tag{11}$$

where P_{size} is zero mean Gaussian function of the Euclidean distance for width w and height h of two detections. Finally, the two-threshold method is applied as follows to obtain the final association result f:

$$f = \begin{cases} 1, & \text{if } K_s \geq \delta_1 \text{ and } I_s \geq \delta_2 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where δ_1 and δ_2 are thresholds for appearance similarity and intersection ratio and they are set to 0.8 and 0.7, respectively.

Then, we use foreground extraction [22] to eliminate false positive detections. If detection is not identified by two-phase association and its foreground area is small, detection is labelled as false positive. If a tracked object is not updated by detection responses for a certain number of frames, then it is eliminated by confirmation-by-classification [21]. Only the reliable associations are passed to the tracking model.

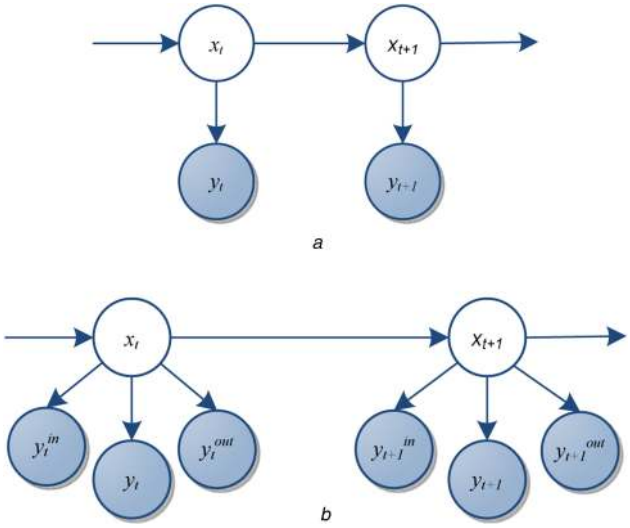

Fig. 3 *Comparison of the proposed model with the standard particle filter*
(a) Standard particle filter model, ***(b)*** Proposed multi observation model

3.3 AdaBoost online learning model

AdaBoost algorithm creates one strong hypothesis from several weak hypotheses [10]. Our weak hypotheses consists of RGB histogram, region covariance and HOG similarity measures:

$$\boldsymbol{h}(A_i, A_j) = \left[cc\left(\boldsymbol{f}_{\mathrm{RGB}_i}, \boldsymbol{f}_{\mathrm{RGB}_j} \right) \quad \sigma(\boldsymbol{C}_i, \boldsymbol{C}_j) \quad cc\left(\boldsymbol{f}_{\mathrm{HOG}_i}, \boldsymbol{f}_{\mathrm{HOG}_j} \right) \right] \tag{13}$$

where cc is the normalised cross correlation coefficient.

By using these weak hypotheses, a strong classifier H is built:

$$H(A_i, A_j) = \sum_{k=1}^{K} \alpha_k \boldsymbol{h}_k(A_i, A_j) \tag{14}$$

where α_k are the weights which are estimated by minimising the loss function. Then we perform binary classification $R_a = \{-1, +1\}$ and calculate the confidence score $\Psi(A_i, A_j)$ between appearance models A_i and A_j as follows:

$$\Psi(A_i, A_j) = \begin{cases} -1, & \text{if } R_a = -1 \\ e^{-cc(\boldsymbol{f}_{\mathrm{RGB}_i}, \boldsymbol{f}_{\mathrm{RGB}_j}) \sigma(\boldsymbol{C}_i, \boldsymbol{C}_j) cc(\boldsymbol{f}_{\mathrm{HOG}_i}, \boldsymbol{f}_{\mathrm{HOG}})}, & \text{if } R_a = +1 \end{cases} \tag{15}$$

Ψ is then used in the state estimation to compare the group members and associate groups in subsequent frames (described in Section 3.4.4).

3.4 Tracking model

In this paper, we refer to an in-group as individuals who belong to a certain group and out-group consists of the individuals that are not part of that group. Unlike the standard particle filter observation model, we use a multi-observation model which consists of in-group (y_t^{in}), out-group (y_t^{out}) observations in addition to a similarity measure (y_t) as shown in Fig. 3. In-group and out-group observations help evaluate the interaction between people and group dynamics following split and merge events.

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

$$P(y_t|x_t) = p(y_t^i, y_t^s|x_t) = p(y_t^i|y_t^s, x_t) p(y_t^s|x_t) \tag{16}$$

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

$$P(y_t|x_t) = p(y_t^i|x_t) p(y_t^s|x_t) \tag{17}$$

The multi-observation model also divides interaction observations into in-group and out-group observations. In-group observation evaluates the interactions between each person in a specific group and other people in the same group while out-group observation performs this evaluation between each person in a specific group and other people outside this group. Since the evaluation of interaction is performed with two disjoint sets of people (in-group and out-group sets), it can be described as follows:

$$p(y_t^i|x_t) = p(y_t^{\mathrm{in}}, y_t^{\mathrm{out}}|x_t) = p(y_t^{\mathrm{in}}|y_t^{\mathrm{out}}, x_t) p(y_t^{\mathrm{out}}, x_t) \tag{18}$$

$$p(y_t^i|x_t) = p(y_t^{\mathrm{in}}|x_t) p(y_t^{\mathrm{out}}, x_t) \tag{19}$$

Thus, the observation model becomes:

$$P(y_t|x_t) = p(y_t^{\mathrm{in}}|x_t) p(y_t^{\mathrm{out}}, x_t) p(y_t^s|x_t) \tag{20}$$

In-group observation $p(y_t^{\mathrm{in}}, x_t)$ is a measure of degree of belonging to a specific group and out-group observation $p(y_t^{\mathrm{out}}, x_t)$ is a measure of degree of not belonging to a specific group.

In the multi-observation model, we use the direction similarity $w_{x,y}^{\theta}$, and normalised spatial distance $w_{x,y}^{d}$:

$$w_{x,y}^{\theta} = \frac{1 + \cos\theta}{2} \tag{21}$$

$$w_{x,y}^{d} = \mathrm{e}^{-\left(d/\min_{k \in s}(k_{\mathrm{w}}, k_{\mathrm{h}})\right)} \tag{22}$$

where θ is the angle between the motion vectors and d is the distance between the individuals x and y, and s is the set of bounding boxes of the person detection result in first frame and the minimum bounding box width k_{w} and height k_{h} in s are used for normalisation. Then, the interaction weight is computed as follows:

$$w_{x,y}^{i} = w_{x,y}^{\theta} w_{x,y}^{d} \tag{23}$$

Defining P^{in} and P^{out} as the sets of persons in in-groups and out-groups respectively, in-group measures $w_{x,y}^{\mathrm{in}}$ and out-group measures $w_{x,y}^{\mathrm{out}}$ between individuals x and y are updated as follows:

$$w_{x,y}^{g} = \begin{cases} 0, & \text{if } w_{x,y}^{d} > w^{d} \text{ or } w_{x,y}^{\theta} > w^{\theta} \\ w_{x,y}^{i}, & \text{otherwise} \end{cases} \tag{24}$$

where w^d and w^θ are the threshold values of spatial closeness and direction similarity and $g \in \{in, out\}$. $w_{x,y}^{\mathrm{in}}$ for an individual is calculated with regards to the persons in P^{in} and $w_{x,y}^{\mathrm{out}}$ is calculated with regards to the persons in P^{out}. Since $w_{x,y}^{d}$ is computed with respect to the normalised distance, we set $w^d = e^{-1}$. In our group definition, people in the same group are assumed to have similar motion directions and the maximum angle between their direction vectors θ is set as $\pi/2$, therefore, w^θ is 0.5.

Since some individuals connect to others indirectly, in-group and out-group weights need to be refined to reflect these indirect connections. 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 $w_{x,y}^{g} = w_{z,y}^{g}$ since x is connected to y through z. After the

refinement of the in-group and out-group weights of all individuals, the event $e \in \{\text{Merge}, \text{Split}, \text{None}\}$ is determined as follows:

$$E(x; y, z) = \begin{cases} \text{Split}, & \text{if } w^{\text{in}}_{x,y} = 0 \\ \text{Merge}, & \text{if } w^{\text{out}}_{x,z} > 0 \\ \text{None}, & \text{otherwise} \end{cases} \quad (25)$$

where $x, y \in P^{\text{in}}$, and $z \in P^{\text{out}}$. Function E detects a split event between individuals x and y, and a merge event between individuals x and z. The individual similarity $p(y^s_t | x_t)$ is calculated using the Euclidian distance.

3.4.2 Motion model: In this model, particle advection and template detection are used together to detect the motion of both sparse and dense groups. Particle advection is mainly used in flow extraction and stability analysis [5, 6] and has been shown to be effective in crowded scenes. On the other hand, template detector is more reliable in sparse environments where individuals are separately discernible. Consequently, the motion model is formulated to combine both approaches and with the object state is composed of the position (p_x and p_y) and velocity (v_x and v_y) of both groups and individuals; $\boldsymbol{x}_t = [p_x\ p_y\ v_x\ v_y]$ as follows:

$$\pi(X_{t+1} | X_{0:t}) = \alpha \pi_{\text{Det}}(X_{t+1} | X_t) + (1 - \alpha) \pi_{\text{PA}}(X_{t+1} | X_{0:t}, y_{0:t+1}) \quad (26)$$

where $\pi_{\text{PA}}(X_{t+1} | X_t)$ and $\pi_{\text{Det}}(X_{t+1} | X_t)$ are hypotheses that generate motion vectors with particle advection and the template detector respectively and α is a parameter that determines the relative weights of these components. α is selected to be inversely proportional to the group density to give higher weight to particle advection in denser regions and higher weight to template detector in sparse regions. We define the group density c_d as:

$$c_d = \frac{n_p}{n_b} \quad (27)$$

where n_p is the number of people in the group and n_b is the number of person rectangles not intersecting with each other in the group. Then, we define α as follows:

$$\alpha = \frac{m}{c_d} \quad (28)$$

where m is the calculated matching measure between the object and the detected template.

As the motion model local-linear dynamics with Gaussian noise is used [3, 4]:

$$\boldsymbol{x}^k_{t+1} = \boldsymbol{B}\boldsymbol{x}^k_t + n_g, \quad \boldsymbol{B} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (29)$$

where $\boldsymbol{B}$ is the particle dynamics matrix, x^k_t is the group state at time t, and n_g is the Gaussian noise.

Since particles in the particle advection model are statically distributed, the following steps are used to eliminate unreliable particles: distance filtering, cross-correlation of particles in successive frames, identifying forward-backward optical flow discrepancies. The distance filter selects the particles in or very close to the tracked object. In cross-correlation of particles on successive frames, these particles are moved by using a forward Lucas–Kanade optical flow [23]. Then, the template match score between moved position and current position for each particle is computed and the particles with a high matching score are kept. For the retained particles, the forward-backward optical flow [24] is applied. In this method, the particles are first moved forward with a forward optical flow, and then moved back with a backward optical flow. The distance between the original position and the new position of the particle is computed as a distance error and only the particles with a very low error are kept. In the final stage, we eliminate unreliable particles and keep the particles having a consistent motion history. This method allows us to identify the right particle in case of an occlusion event. In this method, the velocity vector of each tracked object is retained for a number of frames and the similarity of the current velocity vector is estimated based on the motion history of the tracked object. We only select particles with a consistent velocity vector. Then, the $\pi_{\text{PA}}(X_{t+1} | X_t)$ formula is used to calculate the velocity vector. The template detector model $\pi_{\text{Det}}(X_{t+1} | X_t)$ calculates $[v_x\ v_y]$ estimating the next state of the object through detection. As a template detector, the correlation coefficient is used.

3.4.3 Particle resampling: In standard particle filter model, the number of particles is set once and does not change during tracking. However, since groups are dynamic entities the number of people in a group may change and the number of particles needs to be updated for effective tracking. Particle filter uses sequential importance, evaluating the tracked area as a single-piece and resampling the particles based on the best observation. Instead, we find the best observation for all person regions in the tracked group and resample the particles accordingly. The total number of particles to be used p_t is dynamically computed during tracking as follows:

$$p_t = p \cdot n \quad (30)$$

where p is the number of particles for one person, n is the number of people in the tracked area.

3.4.4 State estimate: State estimation is performed for both groups and individuals. We use the fact that the two groups are the same if and only if number of people in groups is equal and members of the groups are the same. In order to compare the group members and to associate groups and individuals, we use AdaBoost confidence scores Ψ calculated in (15). The state estimation procedure is summarised in Algorithm 1 (see Fig. 4). State estimation returns $S_{\text{est}} = 0$ is the groups are not matched and a state estimate score $S_{\text{est}} > 0$ if they are matched and we select the group with highest S_{est} as the match.

4 Experiments and results

The proposed method was tested on three datasets: friends meet (FM) [3, 4], BIWI [12], and PETS 2009 [25]. The FM dataset consists of 53 sequences including both synthetic and real scenarios. The synthetic set contains 18 simple scenarios (one or two events, two to six individuals), and ten difficult scenarios (multiple events, eight to ten individuals). The real set consists of 15 outdoor sequences where the individuals meet. The queue sequences in this dataset (three synthetic, two real) were excluded from the experiments as in the literature [3, 4]. BIWI consists of two outdoor scenarios and we used six scenarios from PETS.

The proposed method was evaluated in terms of its performance in tracking individuals and groups. False Positive (FP) and False Negative (FN) rates [26] were used for measuring detection performance. Individual tracking performance was measured with ID switch count [27] and the Mean Square Error (MSE) of the estimated positions with their standard deviations. Multi-Object Tracking Precision (MOTP) and multi-object tracking accuracy (MOTA) [28] were used to measure group-tracking performance. Group detection performance was evaluated with the Group Detection Success Rate (GDSR) [3, 4]. We set low-level association thresholds θ_1 and θ_2 to 0.504 and 0.025 respectively.

4.1 Results on synthetic scenarios

Table 1 presents the results of DP2-JIGT [4], DEEPER-JIGT [3] and the proposed method on the FM synthetic where person

Algorithm 1: State Estimate

Inputs: G_1, **Group in the current frame,** G_2, **Group in the previous frame**
Output: $S_{est} \leftarrow 0$, State Estimate Score, initially set to 0
if $size\ of\ G_1 \neq size\ of\ G_2$
// If G_1 *and* G_2 have different number of persons, they are not the same group
return $S_{est} \leftarrow 0$
end if
for *each individual* $p \in G_1$
for *each individual* $q \in G_2$
// compare the group members one by one by using Ψ (calculated using equation (15))
$S \leftarrow \Psi(A_p, A_q)$
if $S > 0$
// accumulate the confidence scores of group members
$S_{est} \leftarrow S_{est} + S$
else
// If there is no match for a person in G_1, they are not the same group
return $S_{est} \leftarrow 0$
end if
end for
end for
// Normalize the score using the number of persons in the group
$S_{est} \leftarrow S_{est}/size\ of\ G_1$

Fig. 4 *Algorithm 1: State estimate*

detector was simulated using detections from the ground truth with a false positive and false negative of 20% and adding Gaussian noise [3, 4].

The proposed framework uses both template detection and particle advection to model the motion of the individuals and groups, and adjusts their relative weights dynamically. In individual tracking, the template detector is expected to have a higher weight; as a result, the performance is adversely affected by detection noise resulting in higher MSE compared with DP2-JIGT and DEEPER-JIGT. However, CIGT outperforms DP2-JIGT and DEEPER-JIGT in group detection and tracking in terms of GDSR, MOTP and MOTA. By eliminating false positives and evaluating the degree of closeness and motion direction with the multi-observation model, the CIGT performs slightly better in terms of 1-FP and GDSR metrics. Furthermore, it better models the group motion and has lower MOTP due to the particle advection.

4.2 Results on real scenarios

FM, BIWI and PETS datasets, each having different properties and challenges, were used to evaluate real scenarios. We used a real person detector [29] for the evaluation of PETS and simulated the person detector for FM and BIWI by generating detections from the ground truth with false positive and false negative rates of 20% and adding spatial Gaussian noise [3, 4]. For group evaluation, we annotated and evaluated ground truth from both group and individual datasets of PETS 2009 [25]. Table 2 presents the results of group evaluation on the real FM dataset.

DP2-JIGT has slightly better performance than the CIGT in group detection since DP2-JIGT performs online modelling for the group. However, the CIGT outperforms DP2-JIGT and DEEPER-JIGT in group-tracking due to the particle advection and multi observation model in the CIGT.

Table 3 presents the results of individual tracking on the real FM dataset. In this case, the intersection degree between ground truth and tracker result is used to calculate MSE and MOTP. In addition, we use re-init (how often the tracker drifts and needs to be re-initialised with detection responses) and ID switch metrics to evaluate individual tracking. Since the CIGT uses the discriminative appearance model [9] to describe the individuals and performs two-phase association to identify them based on their features, the identification based metrics (1-FP and 1-FN) are higher compared with DP2-JIGT and DEEPER-JIGT and thus the CIGT tracker does not need to be reinitialised as often. Furthermore, as a result of two-phase association, the object IDs are mostly correctly identified.

The BIWI dataset consists of two video sequences in which individuals generally walk in one direction and it does not contain

Table 1 Results on the FM synthetic dataset

	MSE[px] (std)	1-FP, %	1-FN, %	GDSR, %	MOTP, px	MOTA, %
CIGT	3.34 (3.80)	**94.79**	90.37	**90.26**	**2.37**	**86.36**
DP2-JIGT [4]	**1.75** (4.76)	93.98	**91.28**	86.91	16.72	71.57
DEEPER-JIGT [3]	2.28 (5.08)	93.12	81.01	78.18	18.16	53.42

Table 2 Results on the real FM dataset

	1-FP, %	1-FN, %	GDSR, %	MOTP, m	MOTA, %
CIGT	97.40	95.81	**95.81**	**0.07**	**94.79**
DP2-JIGT [4]	**97.81**	**97.54**	94.65	0.92	73.85
DEEPER-JIGT [3]	95.72	89.99	85.78	0.87	65.18

Table 3 Results of individual tracking on the real FM dataset

	1-FP, %	1-FN, %	MSE, px	MOTP, px	Re-init, %	ID switch
CIGT	**96.60**	**98.52**	0.21	**0.79**	**0.1**	**14**
DP2-JIGT [4]	81.25	78.11	**0.25**	0.71	3.3	156
DEEPER-JIGT [3]	95.72	89.99	0.24	0.71	3.2	148

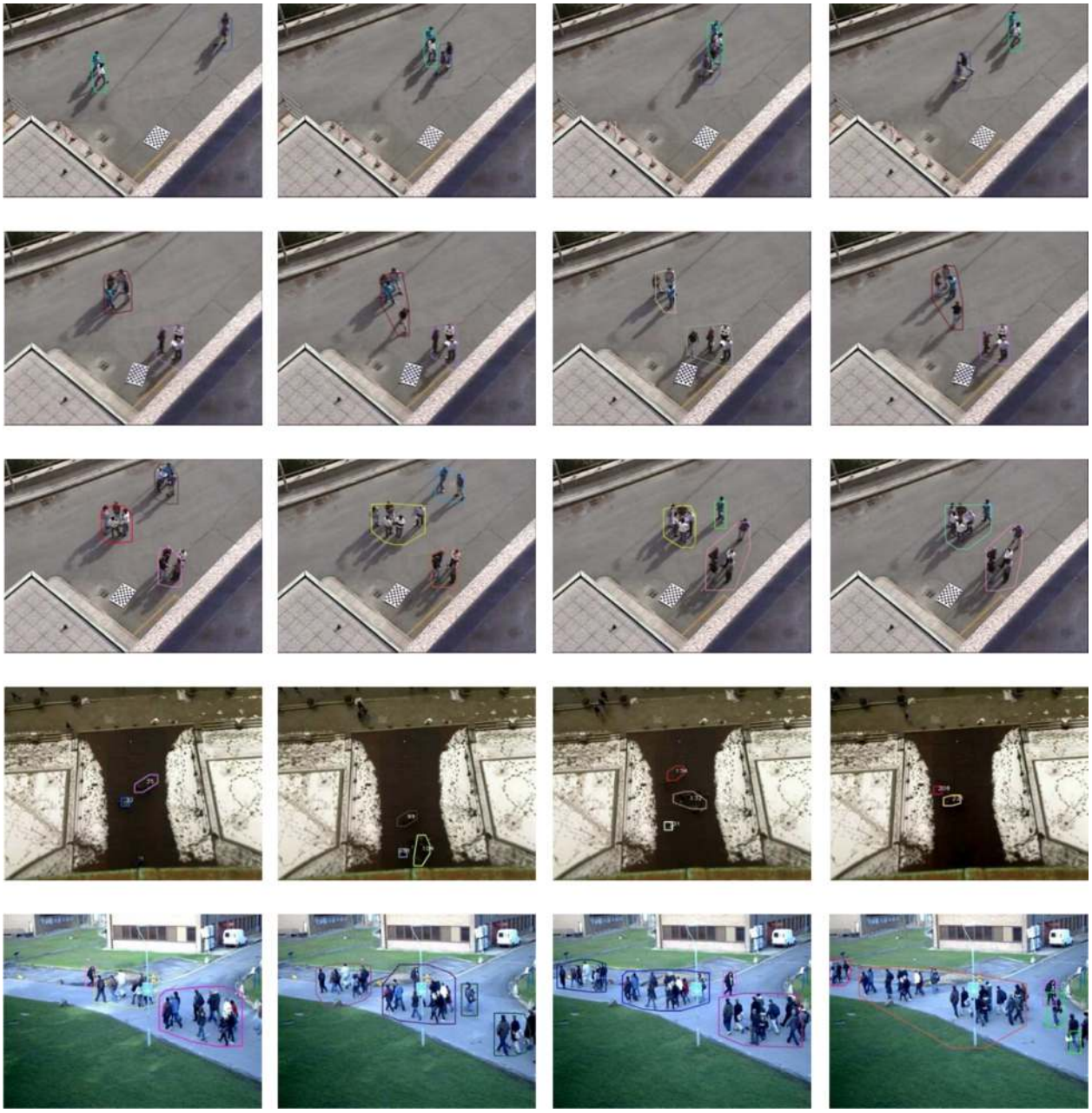

Fig. 5 *Visual results of the CIGT framework on FM (top 3 rows), BIWI (4th row), and PETS (5th row) datasets*

merge and split events. The results on this dataset are shown in Table 4.

All three methods have similar performances in group evaluation on the BIWI dataset. The CIGT outperforms DP2-JIGT and DEEPER-JIGT with much lower false positives. However, bad illumination conditions and small bounding boxes generated by the person detector might reduce the performance of the discriminative appearance model. As a result, some of the valid detections are discarded by false positive elimination increasing the number of false negatives. However, in the CIGT, overall group detection and tracking performance is slightly better; due to advantage of particle advection in denser areas, the CIGT provides better results in modelling the group motion and has a lower MOTP.

The results on the PETS dataset are shown in Table 5. This dataset includes denser groups compared with other datasets different to the other experiments; we used a real person detector [29] to evaluate the tracker performance.

Fig. 5 presents the visual results obtained from the various datasets.

Table 4 Results on the BIWI dataset

	1-FP, %	1-FN, %	GDSR, %	MOTP, m	MOTA, %
CIGT	**91.49**	56.16	**54.57**	**0.31**	**29.66**
DP2-JIGT [4]	37.66	**89.43**	51.86	0.47	22.94
DEEPER-JIGT [3]	53.77	78.00	53.59	0.44	29.43

Table 5 Results on the PETS dataset

	1-FP, %	1-FN, %	GDSR, %	MOTP, m	MOTA, %
CIGT	91.44	71.54	70.63	1.17	49.79

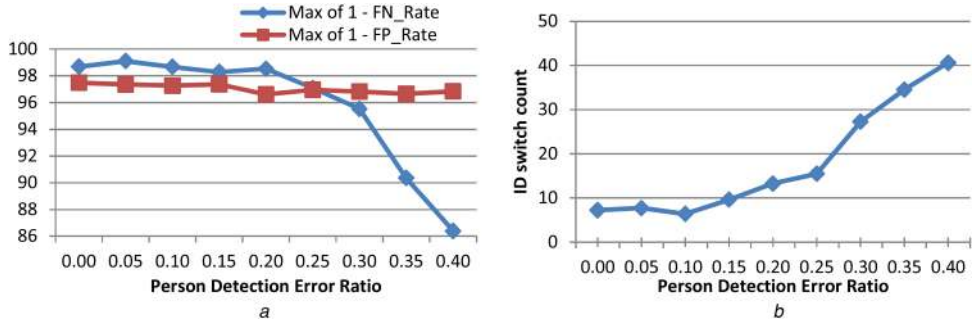

Fig. 6 *Effect of person detection errors on individual tracking in terms of* ***(a)*** False positive and false negative ratios, ***(b)*** ID switch count

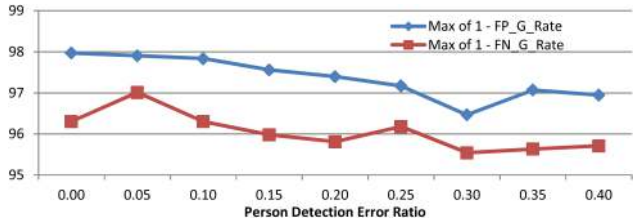

Fig. 7 *Effect of person detection errors on group tracking in terms of false positive and false negative ratios*

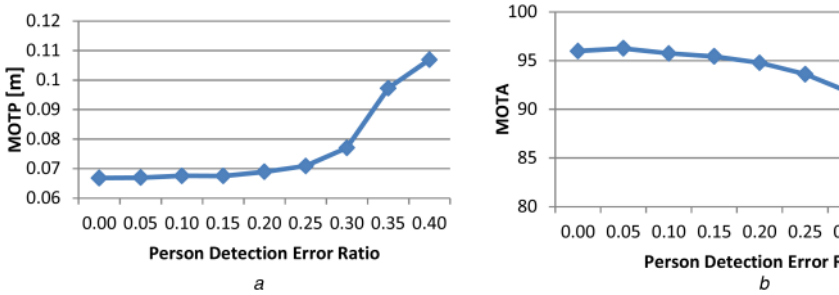

Fig. 8 *Effect of person detection errors in group tracking on* ***(a)*** MOTP in metres, ***(b)*** MOTA

4.3 Person detection error rate and parameter sensitivity analysis

We simulated the person detection by generating 20% false positive and negatives from the ground truth data obtained from previous experiments. In this experiment, we simulated person detections with different error ratios from 0 to 40% in increments of 5% and analysed the effects on both individual and group-tracking. Fig. 6 presents the effects of different person detection error ratios on individual tracking performance.

In individual tracking, although the FN ratio decreases when the error ratio is greater than 0.25, there is no significant change in the FP ratio by the help of false positive elimination. However, the decrease in 1-FN is not proportional to the error ratio and it is still greater than 85% when the error ratio is 0.40. In the literature, the number of ID switches is reported to increase significantly above the 0.25 error ratio. However, the number of ID switches in CIGT did not exhibit any sharp increase with the increasing person detection errors.

As shown in Fig. 7, there is no significant change in the 1-FP and 1-FN rates in group detection. Due to the false positive elimination mechanism of CIGT, the 1-FP rate is affected less compared with the 1-FN rate. These metrics show that the CIGT is robust against person detection errors in group detection.

In group tracking, we evaluate the changes in MOTP [m] and MOTA with respect to the increase in person detection errors as shown in Fig. 8. Since the 1-FN rate in individual tracking decreases, some of the individuals in the group are identified and the central position of the group is drifted. As a result, MOTP increases and MOTA decreases with the increased error rate. However, both MOTP [m] and MOTA are higher compared with similar trackers.

We have also analysed the sensitivity to other parameters on FM synthetic dataset. The experiments showed that varying δ_1 in the range [0.6, 0.9] and δ_2 in the range [0.6, 0.8] had negligible effect on performance. Experiments done by varying θ_1 in the range [0.3, 0.6] show that the performance drops when $\theta_1 < 0.5$ and the change in performance is not significant when $\theta_1 > 0.5$. Varying w^d and w^θ does not affect individual tracking, as expected. On the other hand, for this dataset, group tracking performance increased when w^d and w^θ values increased. While their default values used throughout the paper are sociologically inspired, marginally better results could be achieved by fine tuning these parameters to the specific dataset.

5 Conclusion and future works

This paper proposes a particle filter-based tracker with multi-observation model to track multiple individuals and groups. Different to the standard particle filtering, it provides the flexibility to evaluate multiple observation measures allowing evaluation of social interactions between people and analysis of the group events (merge and split) in group formation. The proposed framework allows groups to be tracked with lower position errors compared with the state-of-the-art methods in the literature. In addition, the discriminative appearance model and online learning provide tracking individuals more accurately and increase the group-tracking performance. In the future, it can be extended with an online learning mechanism for group formation to be used in the multi observation model.

6 Acknowledgment

We are grateful to Dr Loris Bazzani for providing us with the person detection simulation code, dataset parameters and helping with the generation of person detections.

7 References

[1] Andersson, M., Rydell, J.: 'Crowd analysis with target tracking, K-means clustering and hidden Markov models'. Int. Conf. Information Fusion (FUSION), 2012

[2] Chen, T., Schon, T., Ohlsson, H., *et al.*: 'Decentralized particle filter with arbitrary state decomposition', *IEEE Trans. Signal Process.*, 2011, **59**, (2), pp. 465–478

[3] Bazzani, L., Cristani, M., Murino, V.: 'Decentralized particle filter for joint individual-group tracking'. CVPR, 2012
[4] Bazzani, L., Zanotto, M., Cristani, M., *et al.*: 'Joint individual-group modeling for tracking', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, **37**, (4), pp. 746–759
[5] Ali, S., Mubarak, S.: 'A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis'. CVPR, 2007
[6] Solmaz, B., Moore, B.E., Shah, M.: 'Identifying behaviors in crowd scenes using stability analysis for dynamical systems', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, **34**, (10), pp. 2064–2070
[7] Idrees, H., Warner, N., Shah, M.: 'Tracking in dense crowds using prominence and neighborhood motion concurrence', *Image Vis. Comput.*, 2014, **32**, (1), pp. 14–26
[8] Dehghan, A., Idrees, H., Zamir, A.R., *et al.*: 'Automatic detection and tracking of pedestrians in videos with various crowd densities'. Pedestrian and Evacuation Dynamics 2012, 2014, pp. 3–19
[9] Kuo, C.H., Huang, C., Nevetia, R.: 'Multi-target tracking by on-line learned discriminative appearance models'. Computer Vision and Pattern Recognition (CVPR), 2010
[10] Forsyth, D.R.D.: '*Group dynamics*' (Cengage Learning, Wadsworth, 2009)
[11] Helbing, D., Molnar, P.: 'Social force model for pedestrian dynamics', *Phys. Rev. E*, 1998, **51**, (5), pp. 4282–4286
[12] Pellegrini, S., Schindler, K., van Gool, L.: 'You'll never walk alone: modeling social behavior for multi-target tracking'. IEEE Int. Conf. Computer Vision (ICCV), 2009
[13] Tran, K.N., Bedagkar-Gala, A., Kakadiaris, I.A., *et al.*: 'Social cues in group formation and local interactions for collective activity analysis'. VISAPP, 2013
[14] Yigit, A., Temizel, A.: 'Particle filter based conjoint individual-group tracker (CIGT)'. Advanced Video and Signal Based Surveillance (AVSS), 2015
[15] Schapire, R.E., Schapire, R.E.: 'Improved boosting algorithms using con dence-rated predictions', *Mach. Learn.*, 1999, **37**, (3), pp. 297–336
[16] Huang, C., Wu, B., Nevatia, R.: 'Robust object tracking by hierarchical association of detection responses'. ECCV, 2008
[17] Dalal, N., Triggs, B.: 'Histograms of oriented gradients for human detection'. CVPR, 2005
[18] Tuzel, O., Porikli, F.M., Meer, P.: 'Region covariance: a fast descriptor for detection and classification'. Europe Conf. Computer Vision (ECCV), 2006
[19] Wolfgang, F., Moonen, B.: 'A metric for covariance matrices', in (Eds.): '*Geodesy-The Challenge of the 3rd Millennium*' (Springer Berlin Heidelberg, 2003), pp. 299–309
[20] Yang, B., Nevatia, R.: 'An online learned CRF model for multi-target tracking'. CVPR, 2012
[21] Ali, I., Dailey, M.N.: 'Multiple human tracking in high-density crowds'. Advanced Concepts for Intelligent Vision Systems, 2009, pp. 540–549
[22] Yao, J., Odobez, J.-M.: 'Multi-layer background subtraction based on color and texture'. CVPR, 2007
[23] Lucas, B.D., Kanade, T.: 'An iterative image registration technique with an application to stereo vision'. IJCAI, 1981
[24] Kalal, Z., Mikolajczyk, K., Matas, J.: 'Forward-backward error: Automatic detection of tracking failures'. Int. Conf. Pattern Recognition (ICPR), 2010
[25] 'PETS2009 Benchmark Data', 2009. Available at http://www.cvg.rdg.ac.uk/PETS2009/a.html
[26] Smith, K., Gatica-Perez, D., Odobez, J., *et al.*: 'Evaluating multi-object tracking'. Computer Vision and Pattern Recognition (CVPR) Workshops, 2005
[27] Milan, A., Schindler, K., Roth, S.: 'Challenges of ground truth evaluation of multi-target tracking'. Computer Vision and Pattern Recognition Workshops (CVPRW), 2013
[28] Bernardin, K., Stiefelhagen, R.: 'Evaluating multiple object tracking performance: the CLEAR MOT metrics', *J. Image Video Process.*, 2008, **2008**, pp. 1–10
[29] Andriluka, M., Roth, S., Schiele, B.: 'People-tracking-by-detection and people-detection-by-tracking'. CVPR, 2008