# Motion prediction and interaction localisation of people in crowds

Mazzon, Riccardo

# Motion prediction and interaction localisation of people in crowds

by

Riccardo Mazzon

BE in Computer Engineering 2006

MSc in Computer Engineering 2009

A dissertation submitted to

The School of Electronic Engineering and Computer Science

in partial fulfilment of the requirements for the Degree of

Doctor of Philosophy

in the subject of

Electronic Engineering

Queen Mary University of London

Mile End Road

E1 4NS, London, UK

May 2013

*I confirm that the work presented in this thesis is my own and the work of other persons is appropriately acknowledged.*

*Yours sincerely,*

*Riccardo Mazzon*

Thesis supervisor <br> Author

Thesis supervisor
**Professor Andrea Cavallaro**

Author
**Riccardo Mazzon**

**Motion prediction and interaction localisation of people in crowds**

# Abstract

The ability to analyse and predict the movement of people in crowded scenarios can be of fundamental importance for tracking across multiple cameras and interaction localisation. In this thesis, we propose a person re-identification method that takes into account the spatial location of cameras using a plan of the locale and the potential paths people can follow in the unobserved areas. These potential paths are generated using two models. In the first, people's trajectories are constrained to pass through a set of areas of interest (landmarks) in the site. In the second we integrate a goal-driven approach to the Social Force Model (SFM), initially introduced for crowd simulation. SFM models the desire of people to reach specific interest points (goals) in a site, such as exits, shops, seats and meeting points while avoiding walls and barriers. Trajectory propagation creates the possible re-identification candidates, on which association of people across cameras is performed using spatial location of the candidates and appearance features extracted around a person's head. We validate the proposed method in a challenging scenario from London Gatwick airport and compare it to state-of-the-art person re-identification methods.

Moreover, we perform detection and tracking of interacting people in a framework based on SFM that analyses people's trajectories. The method embeds plausible human behaviours to predict interactions in a crowd by iteratively minimising the error between predictions and measurements. We model people approaching a group and restrict the group formation based on the relative velocity of candidate group members. The detected groups are then tracked by linking their centres of interaction over time using a buffered graph-based tracker. We show how the proposed framework outperforms existing group localisation techniques on three publicly available datasets.

# Contents

# Acknowledgements

First, I would like to thank my supervisor Professor Andrea Cavallaro for his continuous support and advice, his uncountable number of suggestions helped me to grow professionally and personally, and in preparation for my future. I would also like to thank my second supervisor, Dr Robert Donnan, and independent assessor, Dr Raul Mondragon, for their extremely useful comments during the various stages of my Ph.D. A big thank you then goes to all past and present people in the AC group, for the enjoyable time spent in the lab together, and for the discussions and support that became fundamental parts of my Ph.D.

I would also express my gratitude to my family that always believed and supported me, even from far away. Finally, a special mention goes to Ilaria, the person who was always the first to celebrate my achievements and give me support in the difficulties encountered during this long journey.

# Previously published work

### Journals

[J1] F. Poiesi, R. Mazzon, and A. Cavallaro. Multi-target tracking on confidence maps: an application to people tracking. *Computer Vision and Image Understanding*, DOI: `http://dx.doi.org/10.1016/j.cviu.2012.08.008`, available online 28 November 2012.

[J2] R. Mazzon and A. Cavallaro. Multi-camera tracking using a Multi-Goal Social Force Model. *Neurocomputing*, Vol. 100, January 2013, pp. 41-50.

[J3] R. Mazzon, S. F. Tahir, and A. Cavallaro. Person re-identification in crowd. *Pattern Recognition Letters*, Vol. 33, Issue 14, 15 October 2012, pp. 1828-1837.

### Conferences

[C1] R. Mazzon, F. Poiesi, and A. Cavallaro. Detection and tracking of groups in crowd. In *Proc. of IEEE Int. Conference on Advanced Video and Signal-Based Surveillance*, Kraków, Poland, 27-30 August 2013.

Electronic preprints are available at:

`http://www.eecs.qmul.ac.uk/~andrea/publications.html`

# Chapter 1

# Introduction

## 1.1 Motivation

With the increasing number of cameras deployed in public areas, it would nowadays be possible to simultaneously monitor a high number of people, if not for the fact that surveillance operators are only able to watch multiple video streams for a limited amount of time due to the repetitive and tedious nature of the task. In order to address this problem, our research aims at an automatic scene understanding where human motion models can generate likely movements for people, and help the prediction of people's motion in unobserved areas and the localisation of interactions.

Surveillance of large areas such as airports and train stations requires the deployment of networks of cameras whose field-of-view (FOV) may be disjointed, thus generating unobserved areas that make the task of tracking a person across the network very challenging (Fig. 1.1). Moreover, the different positioning of the cameras in the network with respect to the scene, involves changes in the pose and scale of people, and changes in illumination that modify the perceived appearance of a person across cameras (Fig. 1.2). Most of the approaches for multi-camera tracking presented in the literature employ machine learning tools to model the expected movements and appearance of people, however these strategies require a training phase performed on datasets that are normally large and time consuming to collate, thus limiting their applicability. In order to avoid extensive learning phases with large amounts of data, we propose a method to propagate people's movements in the unobserved areas that makes use of the knowledge of the map of the area and does not need any training set. People are expected to move towards regions

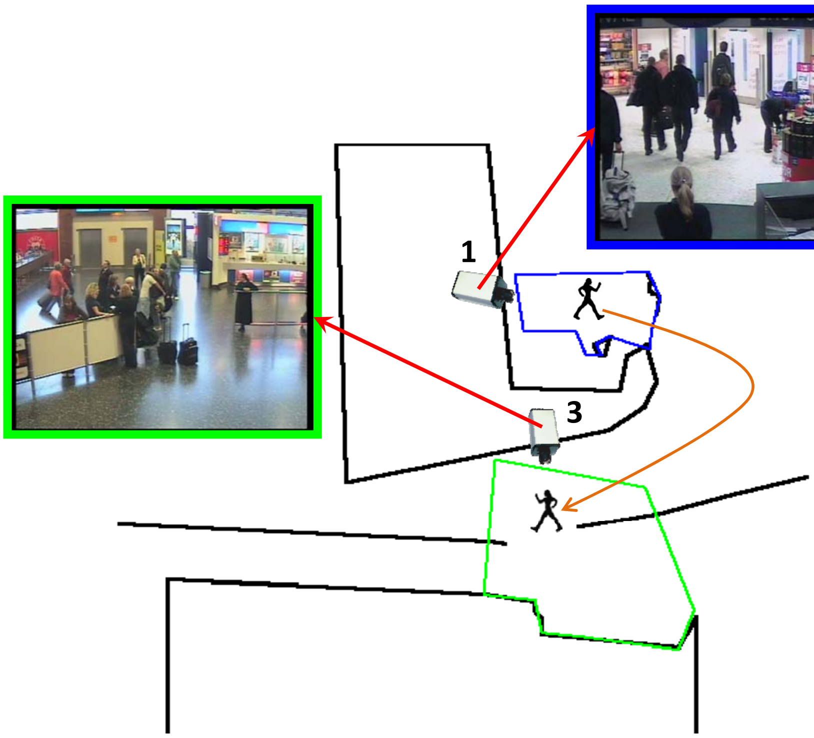


Figure 1.1: Example of person re-identification for multi-camera tracking. Top-view map of the London Gatwick airport (i-LIDS dataset [43]), where the coloured polygons indicate the FOV of Camera 1 (blue) and Camera 3 (green).

of interest and location hypotheses for people's movements are created over time that provide an estimation of where people will be visible again in the next camera. This spatio-temporal information is then merged with appearance cues for the association of people across cameras. In our experiments, we validate the proposed method on video sequences from an airport scenario.

Furthermore, the localisation of group formations is very important to redirect the focus of attention of a surveillance operator towards areas where interactions are happening, for security reasons, and to perform scene analysis, for scene understanding. The task becomes very challenging when people are in a crowd, and when interactions have to be detected instantaneously or within a short period of time in order to make an immediate decision when necessary. In our research, we assume that people's trajectories are known, and we analyse them to extract people's relative velocities and directions of movement over time. Interaction localisation is then performed using a human motion model that generates the expected people's movements in both the situations when people walk alone and in a group. A temporal linking of the centres of interaction allows a clear definition of the movement of each group, if a new group is formed, and if more people join an already existing group. The experiments show that our method is effective

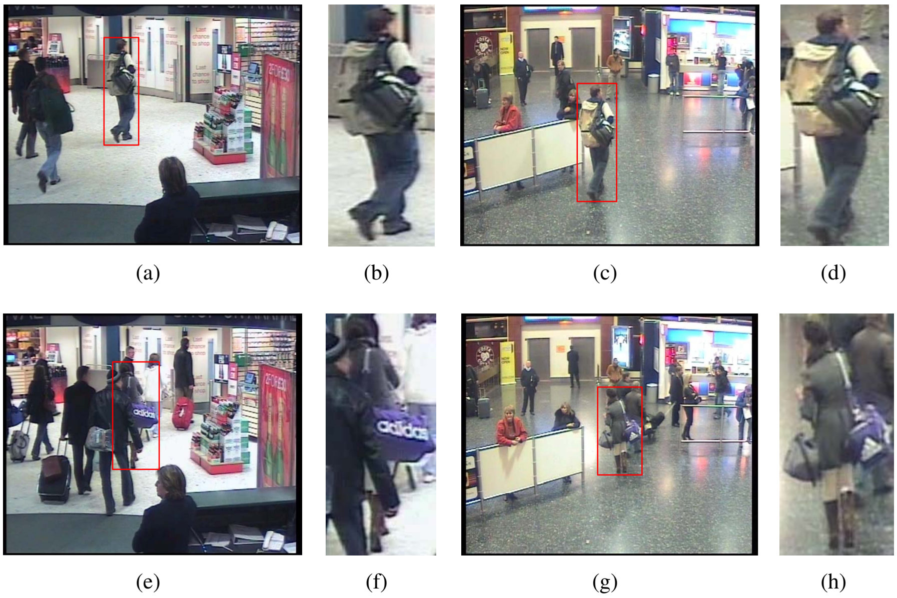


Figure 1.2: Change of people's appearance across cameras [43]. Column 1: camera 1, full frame; Column 2: corresponding crop of a person of interest; Column 3: camera 2, full frame; Column 4: corresponding crop of a person of interest. People appear under different illumination conditions, as shown in (b) and (d), and under different poses and levels of occlusion, as shown in (f) and (h).

in three datasets including a busy square where people are seen to both stand still and walk in a variety of directions.

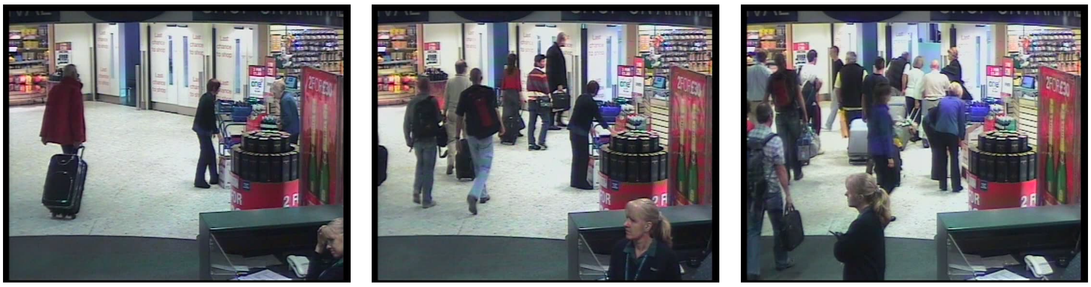
## 1.2 Definitions

The definition of general concepts used in the rest of the thesis is given below:

- *Motion*: The act or process of changing position or place. Motion can be described by temporal features that model the temporal evolution of the video sequence [54].
- *Movement*: An instance of motion [54].
- *Behaviour*: The response of a person to a stimulus or a set of stimuli in a specific context [15]. Behaviour analysis and understanding involve high-level descriptions of motion and common motion patterns in the context where they are estimated.
- *Interaction*: Situation where a person's behaviour is dependent on other people (one or more). Communication needs to be established between the interacting people [15].
- *Group*: A set of interacting people sharing the same objective [15, 78]. Examples are people walking towards the same direction given the same goal to reach, and people standing still talking to each other.

Table 1.1: Level of Service (LoS) [97] and group level for an area of 25 m$^2$. The LoS value is calculated as area over number of people ($N_p$).

| **Level of service (LoS)** | **LoS value** | $\mathbf{N_p}$ | **Group level** |
|---|---|---|---|
| A = free flowing | $> 3.25$ | $< 8$ | Very low |
| B = minor conflicts | 3.25 to 2.32 | 8 to 11 | Low |
| C = some restrictions to speed | 2.32 to 1.39 | 11 to 18 | Moderate |
| D = restricted movement for most | 1.39 to 0.93 | 18 to 27 | High |
| E = restricted movement for all | 0.93 to 0.46 | 27 to 54 | Very High |
| F = shuffling movement for all | $< 0.46$ | $> 54$ | Very very High |



(a) (b) (c)

Figure 1.3: Examples of different crowd density levels in the i-LIDS dataset from London Gatwick airport [43]: (a) Very low; (b) Low; (c) Moderate.

We also define the scene type for a single camera as a combination of *crowd density*, *height* of the camera from the ground and *angle*, $\zeta$, of the camera from the scene. Still [97] proposes the *Level Of Service* (LoS) value as an objective measure for crowd density as it can describe how freely a person moves within a specific area of interest. The LoS value is calculated as the ratio between square meters and number of people, $N_p$, and the LoSs are: free flowing (A), minor conflicts (B), some restrictions to speed (C), restricted movement for most (D), restricted movement for all (E), and shuffling movements for all (F). Given a fixed area of interest, we can also define the *group levels* as a classification of the number of people. Note that a higher number of people corresponds to a lower LoS and to more crowded scenes. Table 1.1 reports LoSs and group levels for an area of 25 m$^2$, and Fig. 1.3 shows an example of very low, low and moderate group levels. In the rest of the thesis, *low-density crowd* corresponds to LoSs A-B and group levels very low-low, *mid-density crowd* to LoSs C-D and group levels moderate-high, and *high-density crowd* to LoSs E-F and group levels very high-very very high. In particular, we concentrate on low- and mid-density crowds since they define the most typical scenes in video surveillance.

Moreover, camera heights from the ground are classified in low, middle (mid), and high positioning. In the case of *low positioning*, the *optical axis* of the camera is at most at head height. In *mid positioning*, the camera is slightly higher than people, up to five meters. Finally, in *high positioning* the camera is well-above the head height (higher than five meters). Each positioning involves different degrees of complexity depending on the application, e.g. cameras at low positioning may facilitate the detection of faces [103], but occlusions are more frequent in the presence of multiple people; while cameras at high positioning reduce the occlusion problem, but the identification of individuals and their body parts becomes very challenging.

Finally, we consider the angle, $\zeta$, formed by the vertical line perpendicular to the ground going through the camera centre and the line going from the camera centre to the scene, being in the range $[0°, 90°]$. An overhead camera that provides a top view of the scene has $\zeta = 0°$, while a camera with horizontal axis has $\zeta = 90°$. In the case of $\zeta = 0°$, the only recognisable part of a person's body is the head, but it is unlikely that it gets occluded; whereas when $\zeta = 90°$ the full body is likely to be visible but more occlusions are expected due to the perspective.

## 1.3 Problem formulation

Let $M$ cameras $C_1, C_2, \ldots, C_M$ with non-overlapping FOVs monitor the area of interest, and let the set of $N$ people $\mathcal{P} = \{P_1, P_2, \ldots, P_N\}$ walk in this area in a low- or mid-density crowd. We define *multi-camera tracking* as the problem of tracking each person in each camera view and associating instances of tracking across cameras. Let us assume $C_1$ to be the first camera where person $P_i$ is visible. A *single-camera detection and tracking* algorithm follows $P_i$ in $C_1$ and creates a trajectory corresponding to the positions of $P_i$ over time. These trajectories can be analysed in order to *localise interactions* among people and define the groups $\Gamma^{\gamma}(t) \subseteq \mathcal{P}$, where $\gamma = 1 \ldots |\Gamma(t)|$, $|\Gamma(t)|$ is the number of groups at time $t$ and people can only belong to one group at the time. Let us now assume that person $P_i$ leaves $C_1$ at time $T_{e_i^1}$ and person $P_r$ appears in $C_2$ at time $T_{s_r^2} > T_{e_i^1}$, where $r = 1, 2, \ldots, N$. The *re-identification* task consists of the association between $P_i$ and $P_r$ using features extracted from single-camera detection and tracking, and information from the scene, environment and camera locations.

## 1.4 Contributions

Given a set of people moving in a crowded environment, our aim is to model these movements for *(i)* motion prediction in unobserved areas and *(ii)* detection and tracking of groups of people. Motion prediction provides candidate positions for people's reappearance using a top view of the monitored site, created from the environment map as a single reference for all cameras. The prediction is performed in the unobserved areas through crowd modelling and with a landmark-based approach. Association across cameras uses spatial locations of the candidates and the appearance of people. Moreover, detection and tracking of groups is performed on single-camera views where the relative velocities of moving and stationary people are employed for interaction localisation.

The main contributions of the thesis are the following:

1. Re-identification algorithm based on motion prediction on a top view. A crowd simulation model is employed for the propagation of people's trajectories from the first camera of the network to the unobserved regions in order to generate candidate locations for people's reappearance. Top view is created using an environment map where people's trajectories from each camera are projected and where spatio-temporal features are extracted. To the best of our knowledge, this is the first application of a crowd simulation model for re-identification [J2].

2. Landmark-based approach on a top view for re-identification, where landmarks correspond to regions of interest in the scene and people's trajectories are propagated in the unobserved regions through the landmarks. The association of people across cameras is performed online using the candidates generated by the motion propagation and a set of appearance features extracted from the upper body [J3].

3. Use of plausible human behaviours for detection and tracking of people interacting in a crowd. Candidates' group members are selected from those that have coherent directions of motion and people approaching a group are explicitly modelled using their relative velocities. The centres of interaction for the detected groups are then tracked with a buffered graph-based tracker that links centre positions over time [C1].

## 1.5 Organisation of the thesis

This report is organised as follows:

*Chapter 1:* Introduction to motion prediction and interaction localisation, description of the fundamental definitions, formulation of the problem and contributions.

*Chapter 2:* Previous work on human motion models and human interaction analysis. Organisation of the state of the art for person re-identification based on features, cross-camera calibration and association. Summary of the main datasets used for validation of the presented works.

*Chapter 3:* Person re-identification based on spatio-temporal candidates generated using human motion models. Association of people across cameras performed using spatial locations of the candidates and appearance of people. Experimental setup and validation for re-identification in two cameras of the i-LIDS dataset from London Gatwick airport.

*Chapter 4:* Group detection using plausible human behaviours. Tracking of the centres of interaction using a graph-based approach. Validation of the approach on BIWI-ETH, BIWI-HOTEL and Student003 datasets.

*Chapter 5:* Summary of the achievements and future work.

# Chapter 2

# Related work

## 2.1 Introduction

In this chapter, we present the state of the art for human motion models, person re-identification, and group detection and tracking. In Sec. 2.2, the human motion models are organised in macroscopic, microscopic and mesoscopic approaches based on how the relationship between individuals is modelled. The literature for person re-identification (Sec. 2.3) is organised into its three main phases, namely feature extraction, cross-camera calibration and association of people [J3]. Group detection and tracking methods (Sec. 2.4) are classified as offline, online, and with latency based on when the decision is taken. Section 2.5 reports a brief description of the main datasets employed for validation and Sec. 2.6 provides a discussion of the presented literature.

## 2.2 Human motion models

We can identify three main strategies for crowd simulation approaches based on how the relationships between pedestrians are modelled, namely macroscopic, microscopic and mesoscopic [119]. *Macroscopic* approaches consider the crowd as an entity, and movements are modelled as a flow that people follow. *Microscopic* approaches consider each person as an entity, and the movement of each person is modelled by taking into consideration interactions among people and the environment. Finally, *mesoscopic* approaches consider groups of people as entities and model their movements by considering the movement of both the crowd as a whole and individuals within the crowd.

Macroscopic approaches are used for person tracking in high-density crowds, where individuals are difficult to be recognised, but the holistic crowd movements can be modelled as a flow. Hughes [42] defines crowds as *thinking* fluids and models them with fluid attributes. An example of application of this method is reported in Bauer *et al.* [7] where a high-density crowd is simulated at the exit of a sport event. Ali and Shah [2] use a similar approach to segment high-density crowds that move towards the same direction in a structured scenario by capturing the underlying dynamics and geometry of the flow. While Rodriguez *et al.* [90] perform people tracking using a Correlated Topic Model (CTM) in unstructured environments where people may have heterogeneous movements. CTM is commonly used to model the correlation of different topics in a document, and in this case the document title is represented by the high-level crowd movement and topics are used to understand the correlation between different motion patterns.

A mesoscopic approach is presented in Ali and Shah [3] where tracking in crowds is performed using cameras placed up high a long distance from the observed scene. People are tracked using *floor fields* in structured environments where high-density crowds have homogeneous flows. Floor fields have three components: *(i)* static floor field to model the scene structure, *(ii)* boundary floor field to model the influence of barriers and walls to the crowd flow, and *(iii)* dynamic floor field to model people's behaviour around the tracked individual. Tracking is performed based on the optimisation of a probabilistic framework of floor fields and a colour patch extracted from each target.

Microscopic approaches are more suitable for modelling and predicting movements of each single person in the crowd. Lerner *et al.* [60] perform crowd simulation by learning people's movement from real sequences using single-camera tracking, thus obtaining realistic crowd behaviours. Brostow and Cipolla [18] spatially and temporally cluster the trajectories of KLT interest points in order to extract their common movements. A discriminator function for motion coherence is then used to count the number of clusters that, in principle, corresponds to the number of people. The algorithm is tested on mid- and high-density crowds where pedestrians' heads or shoulders are the most visible parts with cameras at mid positioning. Another microscopic approach to crowd modelling is the Social Force Model (SFM) first presented by Helbing and Molnar [39] and subsequently refined in Helbing *et al.* [38] by studying *escape panic* behaviours. The SFM is extensively used in crowd simulations and it models forces that guide a person towards a certain goal while avoiding barriers, walls, and other people. Andrade and Fisher [4] simulate two escape scenarios using the SFM, where the crowd simulation is studied in order to understand how people behave in different situations. Results show that the average crowd density increases more in the case of closed exits compared to the case of a collapsed person. SFM is also used for abnormality detection in Mehran *et al.* [72] where the SFM guide the movement of a set of particles spread in the scene and the interaction forces between the agents (in this case particles) are computed using optical flow. Abnormalities are detected by finding uncommon patterns on social interaction forces over time. Furthermore, SFM is applied in single-camera tracking [50, 59, 66, 93]. Johansson *et al.* [50] exploit the SFM in order to understand the forces involved in people's movement. In this case, the parameters for the SFM are learned from a set of tracking results and the model is applied in simple scenarios with overhead cameras where the detection task is already solved. Scovanner and Tappen [93] demonstrate how single-camera tracking can perform better if the motion model follows a minimisation process of social forces, instead of using a linear propagation. Forces due to the environment are not considered since obstacles or walls constraining people's movement are not present in the scene used for validation. Similarly, Luber *et al.* [66] integrate the SFM in a Multi-Hypothesis Tracking (MHT) framework that uses measurements from a laser scanner, and Leal-Teixé *et al.* [59] include it in a graph-based multi-person tracking algorithm where interacting people are also detected.

An alternative solution to the SFM is proposed in Pellegrini *et al.* [80] with the Linear Trajectory Avoidance (LTA) method. In this microscopic approach, the *expected point* where people are likely to move is estimated and a global optimal solution assigns the next step to each target. An improvement of this method is the stochastic LTA (sLTA) [81] where, compared to the original LTA, the final decision is based on a mixture of Gaussians that describes where people are likely to move. A similar approach is presented in Yamaguchi *et al.* [110] where, after learning the model parameters, an efficient energy minimisation algorithm calculates the next step for each person. A different microscopic approach for single-camera tracking is presented in Antonini *et al.* [5], where a Discrete Choice Model (DCM) is the basis of a low complexity tracking algorithm aimed at following people in crowded scenarios. Single pedestrian movements are predicted in the next frame using a discrete grid and the prediction is performed by DCM tuned by a learning phase. In this work, image perspective is relatively easy to rectify since a high-positioning camera is considered.

Moreover, motion models can be applied to predict people's trajectories and goals. Unlike the instantaneous motion models used in single-camera tracking, one of the first attempts of long-term prediction of people's movement towards a goal is presented in Vasquez *et al.* [102]. A Growing Hidden Markov Model (GHMM) is used to predict a target goal after studying its movements by considering the site map divided by a Voronoi diagram, where learning and prediction steps of the GHMM are calculated online using the available observations. Kitani *et al.* [53] propose a Hidden variable Markov Decision Process (hMDP) to estimate the future locations of people by exploiting a training phase where likely paths are learned and by analysing the possible paths a person follows over time. Instead, Gong *et al.* [34] use a motion planning algorithm originally developed for robots to create a set of hypothesis paths people can follow to reach a set of goals by avoiding obstacles. The best hypotheses are then selected by graph search and used to link short-tracklets over time. Finally, Idrees *et al.* [44] integrate collision avoidance, vehicle following, trajectory smoothing and stopping behaviour in a framework that estimates the possible behaviours of cars when unobserved. In this case, the scenario consists of a crossroad controlled by traffic lights where four overhead cameras are directed away from the centre of the crossroad, therefore leaving the crossroad unobserved.

Since we concentrate on low- and mid-density crowds, in this thesis we employ a microscopic motion model for people's movement. In particular, we opt for the SFM that well describes the movement of each person towards a desired goal, and the interaction with the environment and other people. In Ch. 3, the SFM is used for multi-camera tracking to predict people's motion on the map of the monitored area, while in Ch. 4 the SFM is employed to estimate the expected movements of people in order to localise those people that are interacting.

## 2.3 Person re-identification

After performing single-camera detection and tracking [15, 27], person re-identification is used for multi-camera tracking. Unlike previous method-based surveys [26], we organise the re-identification methods into three main phases [J3], namely feature extraction, cross-camera calibration and person association (Fig. 2.1). The first phase is the extraction of *features* from the detected and tracked people. Appearance features include colour, texture and shape [J3]. These features can be extracted from a single snapshot of a target [121] or, when intra-camera tracking information is available, after grouping features over time [12]. Moreover, concatenations of appearance features can also be used [35]. The second phase is *cross-camera calibration*, namely

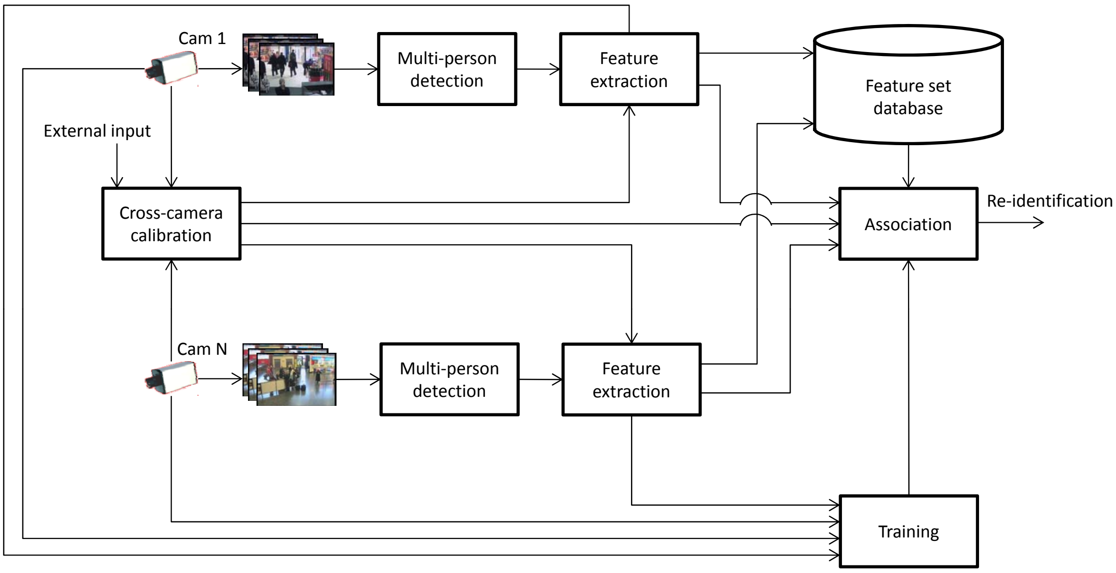


Figure 2.1: Unifying block diagram for person re-identification approaches.

the establishment of the colour and spatio-temporal relationship across cameras that can account for the variability of observations of the same person across different FOVs. Spatio-temporal calibration methods encapsulate information about the camera deployment, the spatial relation between cameras, the entry/exit points in the scene and the travelling time across cameras [47]. Finally, the third phase is the *association* of candidate image regions across cameras to match different instances of the same person using the information extracted in the previous phases. Note that an implicit cross-camera calibration is performed in those approaches that employ a learning process for association [85, 121]. In Sec. 2.3.1, Sec. 2.3.2 and Sec. 2.3.3, we discuss the three phases of person re-identification, respectively, and a summary of the analysed works is reported in Tab. 2.1.

### 2.3.1 Features

Colour, texture and shape are the appearance features commonly used in the state-of-the-art methods for person re-identification. Colour features are extracted from the pixel intensity values of the target; texture features are related to how the various pixels that compose the target are distributed; and shape features are related to the silhouette of the target. In addition, features must be robust to changes in pose since cameras can have different viewpoints. To this aim, features are usually combined together in order to obtain a more representative descriptor of the target. Furthermore, temporal consistency of features can be exploited in order to merge all the

Table 2.1: State-of-the-art methods for person re-identification.

| Ref. | Person representation | | | | Calibration | | Association | | |
|---|---|---|---|---|---|---|---|---|---|
| | Features | | | Temporal grouping | Colour | Spatio-temporal | Measure based | Learning based | Optimisation based |
| | Colour | Texture | Shape | | | | | | |
| [6] | ✓ | ✓ | | ✓ | | | ✓ | | |
| [8] | | ✓ | | | | | | ✓ | |
| [12] | ✓ | ✓ | ✓ | ✓ | | | ✓ | | |
| [16] | ✓ | | | | | ✓ | | | ✓ |
| [21] | ✓ | | | | ✓ | ✓ | | | ✓ |
| [22] | ✓ | ✓ | ✓ | | | ✓ | | | ✓ |
| [23] | ✓ | ✓ | | | ✓ | ✓ | | | ✓ |
| [28] | ✓ | ✓ | | ✓ | | | ✓ | | |
| [32] | ✓ | ✓ | | ✓ | | | ✓ | | |
| [33] | ✓ | | ✓ | | ✓ | ✓ | | | ✓ |
| [35] | ✓ | ✓ | | | | | | ✓ | |
| [36] | | ✓ | | ✓ | | | ✓ | | |
| [40] | ✓ | ✓ | | | | | | ✓ | |
| [45] | | | | | | ✓ | | | ✓ |
| [46] | ✓ | | | | ✓ | ✓ | | | ✓ |
| [47] | ✓ | | | ✓ | ✓ | ✓ | | | ✓ |
| [48] | ✓ | | | | ✓ | | ✓ | | |
| [51] | ✓ | | | | | ✓ | | | ✓ |
| [55] | ✓ | ✓ | ✓ | ✓ | | ✓ | | | ✓ |
| [64] | ✓ | ✓ | | | ✓ | | ✓ | | |
| [65] | ✓ | | | | | ✓ | ✓ | | |
| [69] | ✓ | | | ✓ | | | ✓ | | |
| [76] | ✓ | ✓ | | | | | ✓ | | |
| [83] | ✓ | | | | ✓ | | | | ✓ |
| [84] | ✓ | | | | ✓ | ✓ | ✓ | | |
| [85] | ✓ | ✓ | | | | | | ✓ | |
| [92] | ✓ | | ✓ | | ✓ | | | | ✓ |
| [94] | ✓ | | | | ✓ | | ✓ | | |
| [99] | | ✓ | | ✓ | | | | ✓ | |
| [106] | ✓ | ✓ | ✓ | | | | | ✓ | |
| [121] | ✓ | ✓ | | | | | | ✓ | |

available information taken from multiple snapshots (person patches) over time.

*Colour* is the most commonly used appearance feature encoded in the form of either histograms [21, 22, 28, 32, 33, 35, 47, 55, 65, 76, 85, 92, 94, 121] or cumulative histograms [12], which are simple to compute and scale invariant. Different colour channels and their combinations can be used: the Hue channel from the HSV colour space [76]; the Hue and Saturation channels jointly [32]; the three channels of the HSV colour space [28, 40, 92]; or the Lab colour space [40]. Also, the histogram of the RGB colour space is widely used [12, 21, 22, 23, 33, 47, 65, 84, 94]. However, Consensus-Colour Conversion of Munsell (CCCM) has proved to be a better colour space compared to RGB and HSL[1] [16]. A concatenation of histograms from RGB, YCbCr, and HS (from HSV) colour channels (Fig. 2.2) is adopted in Gray and Tao [35], Prosser *et al.* [85] and Zheng *et al.* [121]. An analysis by boosting classifier [35] shows how, for the re-identification task, the Hue channel is the most discriminative followed by Saturation, Blue, Red, and Green channels. However this analysis is limited to scenes where people are fully vis-

[1] HSL is a cylindrical-coordinate representation colour space similar to HSV.

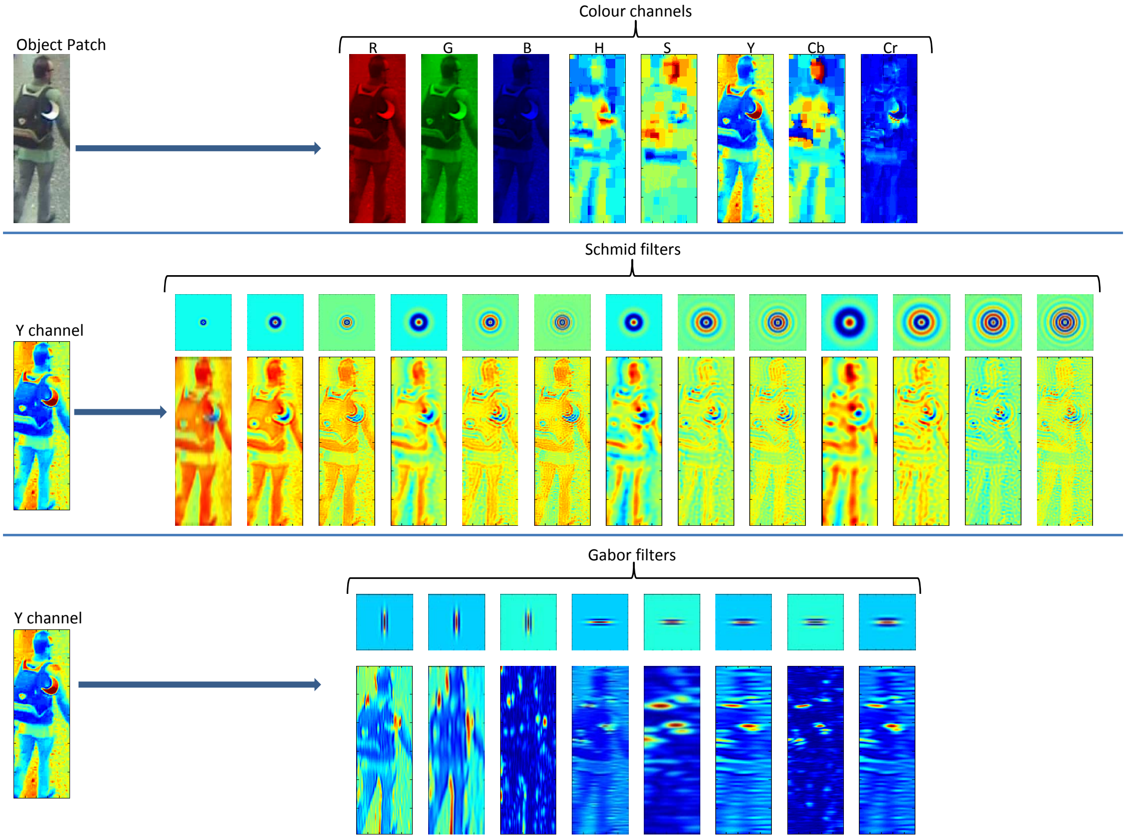


Figure 2.2: Example of colour and texture features extraction. Colour features can be extracted from different channels (top row). Textural features can be extracted by applying Schmid (middle row) and Gabor (bottom row) filters on the Y channel [35].

ible. Alternatively, the two chrominance channels from the YUV space are used in [48], where a Gaussian Mixture Model is applied to find the most relevant colour clusters, whose centres are adopted as descriptors. The Dominant Colour Descriptor (DCD) [6] and the Major Colour Spectrum Histogram Representation (MCSHR) [69] compute the most recurrent RGB colour values that are then used to represent a patch. Moreover, Maximally Stable Colour Regions (MSCR) [28] extracts the homogeneous colour in the person patch by grouping neighbouring colour blobs. Finally, camera parameters and reflectance of the objects' surface can be studied to obtain the main appearance characteristic of the target [47]. DCD, MCSHR, MSCR and object reflectance are features applicable only when a person is captured at medium/high resolution (i.e. larger than $100 \times 40$ pixels) and there is full-body visibility [J3].

The spatial distribution of the intensities in a person patch can be a key feature for person re-identification. Gabor and Schmid filters (Fig. 2.2) define two kernels for *texture* extraction applied to the luminance channel [35, 85, 121]. Gabor filters are linear filters similar to the way the human visual system is expected to describe horizontal and vertical structures, while Schmid filters are rotational invariant Gabor-like filters. HAAR-like features can be used to extract relevant textural information from the person patch with the aim of finding recurrent colour distributions [6]. Furthermore, the *ratios* between different regions in a patch can be used as a discriminative feature. Ratios of colours, ratios of oriented gradients and ratios of saliency maps can also be used as textural features [12]. Similarly, Recurrent High-Structured Patches (RHSP) extract the most common blobs in the person patch [28]; in addition to this, salient spatio-temporal edges (edgels) obtained from watershed segmentation carry information of the dominant boundary and of ratios between RGB channels [32]. The distribution of spatial patches can be directly extracted in the frequency domain where Discrete Cosine Transform (DCT) coefficients can be used as textural features [8]. Finally, spatial patch distribution can be extracted by computing the first and the second derivatives of the person patch resulting in a covariance matrix [55, 106]. Symmetric regions of a patch can be an alternative to the covariance matrix. The symmetry within the person patch can also be exploited in the extraction of local features, weighting each feature based on their position with respect to the symmetric part [28]. These filtering methods are robust to illumination changes but cannot deal with large pose changes. In particular, Gabor and Schmid filters, and HAAR-like features are local descriptors suitable for small patches, while the ratios, RHSP, salient edgels, DCT coefficients, and covariance matrix can only be applied to people's patches at medium/high resolution. Furthermore, a Histogram of Oriented Gradients (HOG) gives information on the orientation of the edges in the patch [55, 106], creating a feature that models the shape of the object by its edge distribution. However, HOG features are only invariant to changes in illumination, and not to changes in pose and scale. A possible solution is the CI_DLBP, a combination of Colour Intensity (CI) and Distance based Local Binary Pattern (DLBP) calculated in different colour spaces and their variations, and extracted from upper and lower body parts [64]. Alternatively, a combination of HSV and Lab, as colour features, with Local Binary Patterns, as texture features, extracted from small rectangular regions of the full body can be employed [40]. The *silhouette* of a person has also been used when cameras are geometrically calibrated. The bounding box around each person that comes from single-camera tracking can be exploited by extracting the angle formed by the vertical edge and the diagonal of the bounding box [22, 33]. A more general feature is the height of the target when calibration information is available [12, 92]. Finally, *interest points* can be used for re-identification in the case of variations in scale, pose and illumination [9]. Examples are SIFT [99], SURF-like features [36, 76] and the Hessian Affine invariant operator [32].

When intra-camera tracking information is available, features extracted from single images can be grouped over time either by *temporal* accumulation [36] or by clustering [28]. Then, the most representative patch of the set is kept as representative for the specific person. A spatio-temporal over-segmentation of patches over 10 frames can be used to create a signature for each person [32]. However, the most common approach is to keep all the available features extracted from single patches over time and then perform association by analysing the similarity among all the available features [12, 47, 55]. Features can also be incrementally updated over time, for example using Incremental MCSHR (IMCSHR) that updates MCSHR in order to increase robustness in situations where there are abrupt changes in illumination [69]. Finally, features extracted from patches of the same person over time can be used as a set of positive samples for training a learning based method [6]. In general, using temporal information, the effects of light variations within the same camera and short occlusions of people are reduced because more representative features for each target are created.

### 2.3.2 Cross-camera calibration

Cross-camera calibration includes colour calibration and spatio-temporal calibration. Different illumination conditions across cameras can be compensated with robust features, as discussed in Sec. 2.3.1, and via colour calibration where cross-camera *colour calibration* models the colour relationship between camera pairs [83]. This approach requires a learning stage where, for each camera, a relationship must be found and updated over time to cope with daily changes in the lighting conditions. Examples of colour calibration include the Brightness Transfer Function (BTF) [47] and the Colour Transfer Function [48]. It is demonstrated that all BTFs lie in a low dimensional space that is discovered using Principal Component Analysis (PCA) on RGB colour intensities [47]. In this case, colour calibration is based on a linear function. Possible improvements of BTF are the Cumulative BTF (CBTF) where the contribution of less common training samples is taken into account [64, 84], the unsupervised incremental CBTF [94], and the work of Chen *et al.* [21] that learns offline the BTF for each camera pair and then updates it over time with an incremental Probabilistic PCA. Clustering with GMM can be used on RBG colour space [21, 33] or, in order to find an affine colour calibration transformation, on the chromaticity space [23, 48]. An alternative approach [92] calculates the best value to be multiplied to the RGB intensities. Colour calibration can perform well in the case of large inter-camera illumination changes, however it can only be applied to scenes where abrupt illumination changes are unlikely to happen.

The knowledge of the environment in which the cameras are deployed can be used to restrict the re-identification task within a certain time interval and certain regions of the monitored scenario, by estimating when and where people are going to reappear in the next camera (*spatio-temporal calibration*). The average travelling time across cameras and the expected entry/exit points in the scene can be learned [21, 33, 47, 55, 71], or manually selected [23, 84]. Learning the time it takes to travel across cameras can be complemented by the learning of probable entry/exit regions in the camera network [47, 71]. However, when the relative camera positions are known, people's location and speed can be discriminative features for each person [22]. The main limitation of these approaches is that they are only suitable for scenarios where unobserved regions are easy to model and people always follow the most common paths. Instead of modelling entry/exit regions and average travelling time for people, a possible solution is to learn the activities in each camera and then obtain spatio-temporal information for person re-identification by cross-camera activity matching [65]. Alternatively, in our approach we propagate people's movements in the unobserved regions in order to create a set of potential locations for people's reappearance in the next camera. We propose two models. The first model discretises the unobserved regions with a set of landmarks that the hypotheses for people's movements must traverse [J3] and the second predicts people's movement with a SFM-based method by assuming that people move towards a set of regions of interest, and avoid walls and barriers [J2].

### 2.3.3 Association

The core of a person re-identification method is the definition of how to match features of candidate people. In order to associate the same person across cameras, we can measure the feature (dis)similarity, use a trained classifier, or perform an optimisation process. The most direct and straightforward approach is to compute the distance or the correlation between feature sets. Lower the distance, higher the similarity (lower the dissimilarity), and opposite for correlation. Due to the challenging nature of the problem, distance and correlation measures are not always sufficient to obtain good association results. More robust approaches involve learning based methods. With these methods, the association problem is converted into a class-based problem where the similarity between features is defined using a training phase. Finally, there exist different approaches that tackle the association problem as an optimisation problem. This optimisation based person association is a maximisation/minimisation of a probability or of an energy-based framework. The framework is composed of all the features from people in the scene, matched using measures and/or learning based methods.

Person association using direct measures estimates the point-to-point *dissimilarity* between feature vectors. The Euclidean distance is used for vectors representing colour values [23, 28], interest points, or hypotheses about a person's location [32, J3]. The Euclidean distance between two colours is also included in an ad-hoc similarity measure created to compare two DCD feature sets [6]. Alternative measures are the sum of quadratic distances [76] and the sum of absolute differences [36]. The main disadvantage of these point-to-point distance measures is that the single elements of the feature set are considered separately and the holistic information of the set is neglected. Other distance measures include the Kullback-Leibler Distance (KLD) [12, 48] and the Bhattacharyya Distance (BD) [28, 65, 84, 94]. KLD is a directed non-symmetric divergence measure used on feature sets composed of histograms. An additional measure derived from the Kolmogrov distance is introduced in Madden *et al.* [69] to compare IMCSHR features, while Lian *et al.* [64] employ a Chi square distance applied to different body parts. Correlation between colour histograms and HOGs of the objects is used in Kuo *et al.* [55]. In these methods, the most challenging part is the selection of the best distance for the specific set of features which is usually chosen by trial and error.

Approaches based on measuring similarity between feature sets are not robust to illumination changes unless cross-camera colour calibration is performed, as discussed in Sec. 2.3.2. As an alternative, a *classifier* can be trained to learn the changes between cameras using labelled features. Support Vector Machines (SVM) can be employed with DCT features [8] and SIFT [99]. An improvement of the standard SVM is the Ensemble SVM, as it reduces the computational cost of rankSVM for high-dimensional feature spaces as well as converting the re-identification problem into a ranking problem [85]. Furthermore, AdaBoost is applied to person re-identification to learn weak classifiers based on different feature sets and to identify the most discriminative features [35]. A different learning-based approach is based on Probabilistic Relative Distance Comparison (PRDC) [121]. PRDC maximises the probability of correct matches while minimising that of wrong matches by learning the best distance measure for the association. A similar approach but computationally more efficient is the one presented in Hirzer *et al.* [40] where a

Mahalanobis distance is efficiently learned after PCA has been applied to the feature space. Unlike direct distances, these methods are less sensitive to feature selection because the importance of each feature for association is learned using a training set. However, good performances are only expected in those cases where the testing set is similar to the training set (e.g. training and testing sets are extracted from the same camera network). A brute-force solution would train the classifier in each scenario using an ad-hoc training set, but this may not always be feasible as the labelling of training sets is normally time consuming. In order to address these limitations, transfer learning techniques may be employed [62]. These techniques can transfer the knowledge learned in one domain (from a training set) to a different domain, thus limiting and/or avoiding the necessity of a new training set specifically created for each case[2].

Other approaches use *optimisation*-based algorithms. One of the first attempts is a Bayesian formulation and MAP [51] where the implementation is based on Linear Programming. This method was further improved [33, 46] using a learning approach for colour calibration and Parzen window for spatio-temporal calibration. The concept of belief/uncertainty assignment can be exploited and the decision for the association problem can be made on specific ad-hoc rules [22]. Also Bowden and KaewTraKulPong [16] designed specific ad-hoc rules for person's reappearance that are employed in a probabilistic framework. Euclidean distance minimisation can be used between trajectories projected to the extension of the FOV lines of the cameras [45]. An alternative approach finds the maximum likelihood Probability Density Functions (PDF) of appearance and spatio-temporal features of different observations of the same object using a weighted sum optimisation [21] or a split graph [47]. Re-identification can also be performed by Hungarian algorithm using colour, texture, and spatio-temporal features [55], where the 'potentially' correct matches are selected by Multi Instance Learning (MIL) boosting on the spatio-temporal features. A similar approach is the Multiple Component Matching (MCM) [92] where only positive samples are used for training, and Hausdorff Distance and BD are used as measures. Finally, dynamic programming is used to find the fitting of body models across cameras [32]. The main drawback of optimisation-based approaches is that they operate in a batch mode and cannot be run online.

When analysing re-identification algorithms using the ranking score assigned to each person, results on methods solely based on appearance usually achieve less than 40-50% [121] for the

[2] For the interested reader, a survey on transfer learning techniques can be found in Pan and Yang [77].

first ranking position (the real re-identification score) when 476 images of 119 people are considered. Re-identification algorithms that operate in batch mode and also exploit spatio-temporal features, can achieve results over 90% [47] for the first ranking position in scenes where there are on average 42 transitions across cameras, linear motion of people in unobserved regions and full-body visibility. Nevertheless, methods solely based on appearance can be tested using single snapshots of people (Sec. 2.5.1) and they become very important when cameras are located far apart. In this scenario, spatio-temporal calibration is very challenging and spatio-temporal features become less reliable. However, the propagation of people's motion [J2, J3] can be applied to a camera network with non-trivial layout where a set of spatio-temporal cues enhances the performance of appearance features for association.

## 2.4 Human interaction analysis

The localisation of people's interactions can be performed online, offline or with latency based on when the methods provide an output with respect to the dynamics in the scene. *Online* methods enable the localisation of interactions without using future information [10, 20, 118]. Bazzani *et al.* [10] propose a Decentralised Particle Filtering (DPF) for group detection and tracking where the states of the filter contain the position and velocity information of people, and labels of the group affiliation of each person. Furthermore, Zanotto *et al.* [118] employ an unsupervised method for group detection based on Dirichlet Process Mixture Model (DPMM) with real-time processing, where motion patterns along with social constraints based on rules of proxemics are used to determine group formations. Real-time processing is also achieved in Chang *et al.* [20] where soft grouping structures are detected using a pairwise measure on people's motion and group connectivity through graph-cut methods. The soft group structures are then analysed to estimate the specific grouping scenario.

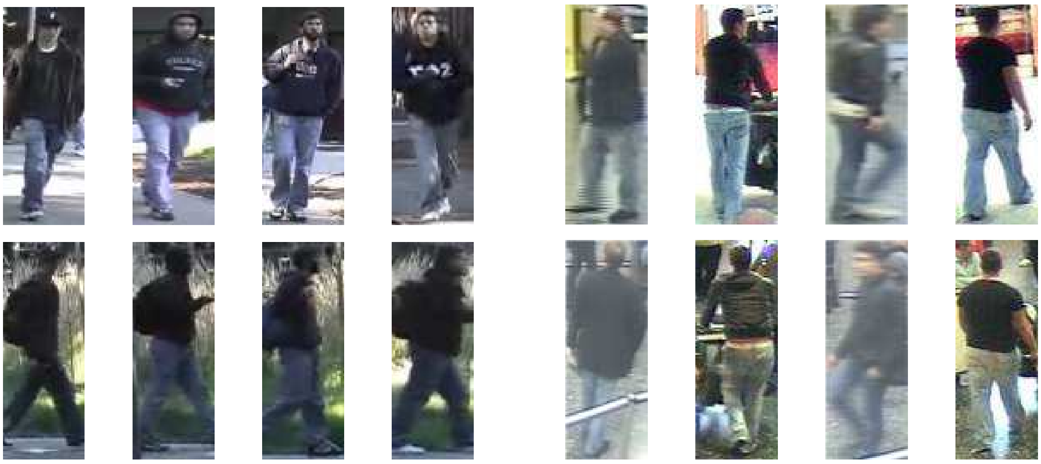
*Offline* methods process the information extracted from the whole video in a batch. The extracted human motion patterns (e.g. position and velocity) are temporally analysed in order to determine the affiliation of each subject to a particular group and group detection is performed on the overall permanence of people within a group. Common directions and velocities of humans processed with a bottom-up hierarchical clustering [30, 31], an optimisation based on Lagrangian theory [86], an SVM classification [110] or a hypothesis testing scheme [117] can be employed to detect groups. The clustering proposed in Ge *et al.* [30, 31] uses the symmetric Hausdorff measure to iteratively evaluate the distance between people belonging to different groups; the Lagrangian theory proposed in Qin and Shelton [86] uses an iterative two-step algorithm that first links tracklets by Hungarian algorithm and then groups of people are generated using K-mean clustering; an extensive training on possible distances and velocities are used in the SVM approach in Yamaguchi *et al.* [110]; while Yücel *et al.* [117] employ a hypothesis testing scheme where positions are modelled with Minimum Spanning Trees and directions with von Mises distributions. Alternatively, techniques based on *social forces* can be used to model behaviours through the analysis of relative motion patterns [59, 74, 75, 105]. The experimental study of how people self-organise themselves when walking in crowded environments reported in Moussaïd *et al.* [74] is used in Moussaïd *et al.* [75] to embed group forces in the SFM. Šochman and Hogg [105] exploit this modification in an error-minimisation framework to detect groups in mid-density crowds. Finally, Leal-Taixé *et al.* [59] detect interacting people using the SFM where people's proximity is analysed over time. These approaches aim to recognise groups that contain people who know each other, rather than localising short interactions.

Methods *with latency* can use the SFM for group modelling, whereas the final group decision [105] is taken with an offline error minimisation process. Interestingly, this algorithm outperforms state-of-the-art approaches in several difficult scenarios even if the group modelling only analyses people's movements and does not employ explicit human behaviour constraints for group formation. However, this approach struggles to detect instantaneous interactions in some simple situations which can be partially addressed by offline approaches [31]. In order to address these situations while maintaining the method with latency, we include in the group modelling two human behaviour constraints that consider relative directions and velocities of people, and are calculated instantaneously [C1]. In particular, only people walking together can be detected as part of the same group and only people stopping in proximity of a static group can be detected as belonging to that group.

## 2.5 Datasets

### 2.5.1 Person re-identification

Existing person re-identification methods are validated on snapshot-based or video-based datasets. On the one hand, VIPeR [28, 35, 64, 85, 92, 121] and i-LIDS-static [6, 28, 85, 121] are the most common snapshot-based datasets used to validate appearance-based methods mostly con-



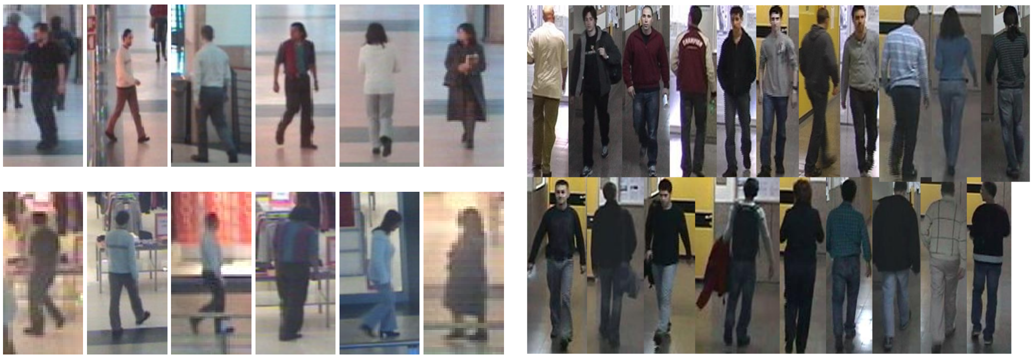
(a) VIPeR

(b) i-LIDS from London Gatwick airport

Figure 2.3: Examples of different snapshot-based datasets used to test re-identification methods [85].

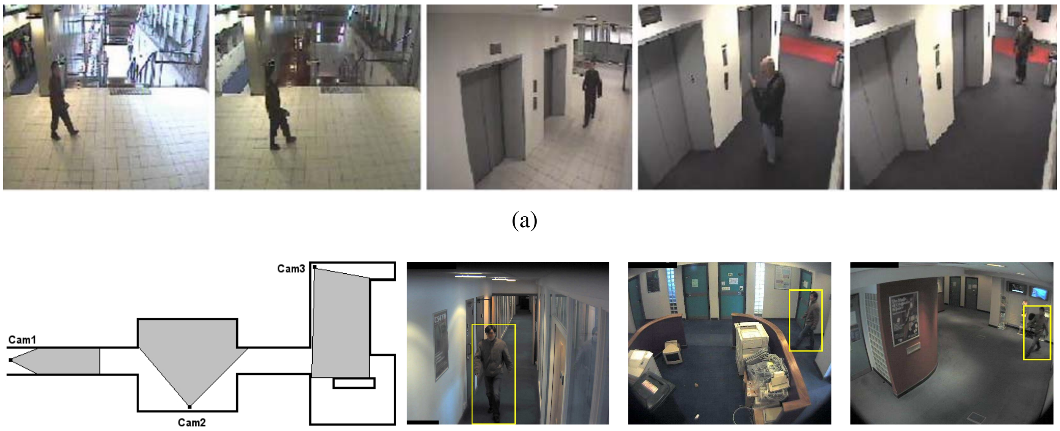
taining people with full-body visibility. VIPeR consists of 632 images taken from two outdoor views [35], while i-LIDS-static contains from 44 [6] to 476 [121] images of people taken from four cameras at London Gatwick airport. Since motion information is not available, only methods based on appearance can be tested on these datasets. Figure 2.3 reports examples of these datasets.

On the other hand, video-based datasets are more representative of video surveillance camera networks, but no standard datasets have been used in literature. Datasets with only one person or only isolated people walking in the scene are CAVIAR [6, 36, 76, 99], *Terrascope* [48], GBSEO [12], PETS2007 [94] and others self-made [21, 22, 32, 33, 47, 64, 69, 84, 106]. CAVIAR is recorded in a shopping centre in Portugal and it has two overlapping cameras with large illumination changes (Fig. 2.4(a)). *Terrascope* has nine indoor cameras with overlapping and non-overlapping FOVs that cover a wide area, where eight people walk and act in an office environment. GBSEO has two non-overlapping indoor cameras (Fig. 2.4(b)). PETS2007 has four cameras from Glasgow airport, UK. Furthermore, three indoor cameras are used in Madden *et al.* [69] (Fig. 2.5(a)) and in Prosser *et al.* [84] (Fig. 2.5(b)). In Cheng *et al.* [22] a dataset with four outdoor cameras is used (Fig.2.6(a)), while three outdoor cameras are used in both Gheissari *et al.* [32] and Wang *et al.* [106] (Fig. 2.6(b)). Javed *et al.* [47] employ a video-based dataset with three sequences composed by up to three cameras from indoor and outdoor scenarios with large illumination changes and up to four fully visible people (Fig. 2.6(c)). Although these datasets



(a) (b)

Figure 2.4: Examples of different video-based datasets used to test re-identification methods. (a) CAVIAR [6]; (b) GBSEO [12].

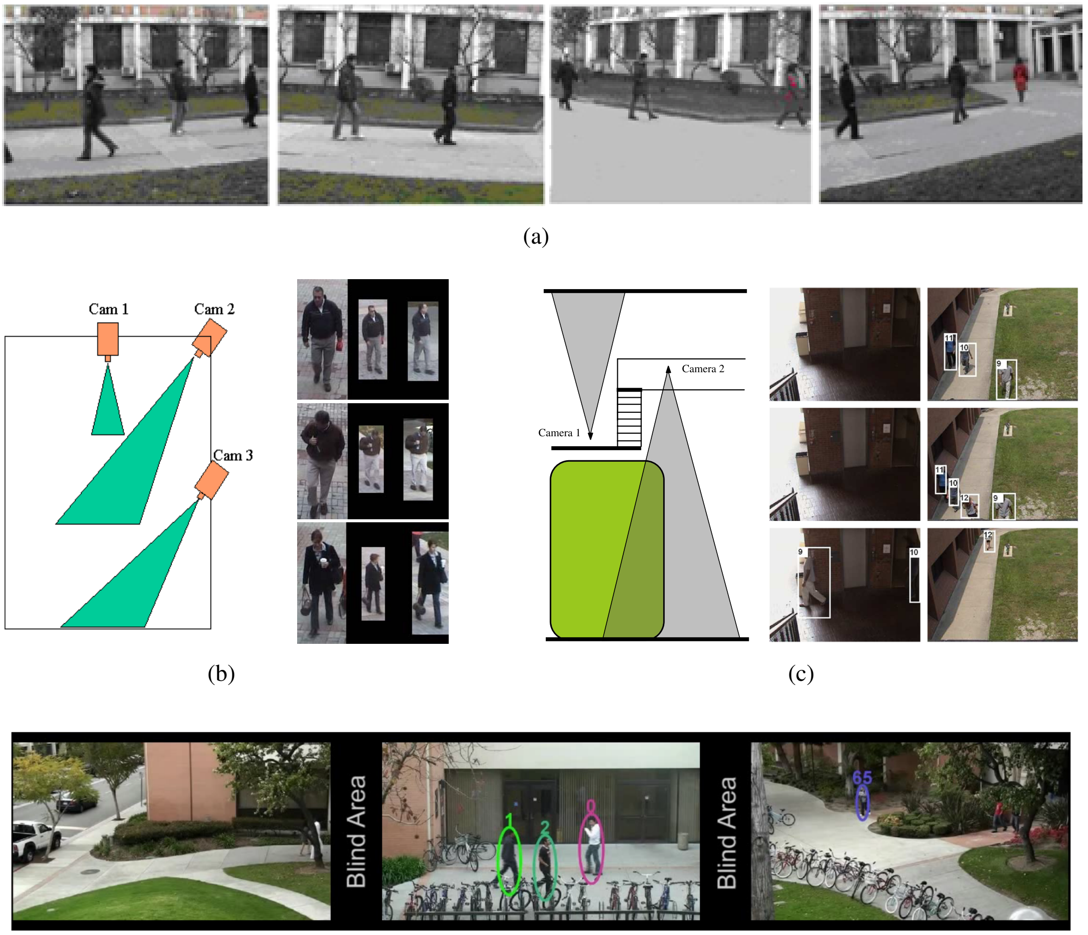


(a)

(b)

Figure 2.5: Examples of different video-based datasets used to test re-identification methods. (a) Madden *et al.* [69]; (b) Prosser *et al.* [84].

have been designed to validate re-identification methods, they may not be completely representative of common scenarios as people in the scene are never occluded and methods based on appearance can always perform feature extraction on the full body of a person. A more challenging dataset in terms of occlusions is presented in Kuo *et al.* [55] and composed of three outdoor cameras at mid-positioning where up to ten people walk alone or in small groups (low- and mid-crowd density) (Fig. 2.6(d)). We can consider this last dataset as more representative of real scenarios, even if the camera deployment is quite simple as cameras are placed next to each other with similar point of views. Finally, Colombo *et al.* [23] present a dataset composed by 29 cameras from different stations of the Turin underground (Fig. 2.7), where fourteen people travelling in the camera network are annotated.



(a)

(b)

(c)

(d)

Figure 2.6: Examples of different video-based datasets used to test re-identification methods. (a) Cheng *et al.* [22]; (b) Gheissari *et al.* [32]; (c) Javed *et al.* [47]; (d) Kuo *et al.* [55].

### 2.5.2 Human interaction analysis

The validation of group detection and tracking methods can be performed in self-made datasets used for testing specific characteristic of the methods or in datasets recorded in public scenarios. Examples of self-made datasets are the ones proposed in Moussaïd *et al.* [74, 75] used to understand social interactions in a crowd and the Friends Meet (FM) [10] (Fig. 2.8(a)) where people stand still in well-defined small groups or join already existing groups.

The most popular datasets from public scenarios are BIWI-ETH [10, 31, 80, 105, 110, 118, C1] (Fig. 2.8(b)) recorded at a University entrance, BIWI-HOTEL [31, 105, 110, 118, C1] (Fig.2.8(c)) that recorded a pavement next to a tram stop with a overhead camera, and Student003 [60, 105, 110, 118, C1] (Fig. 2.8(d)) obtained from a busy square. Other datasets include Zara01 and Zara02 [60, 110, 118] (Fig. 2.8(e)) that show a pavement in front of a shop; Town-

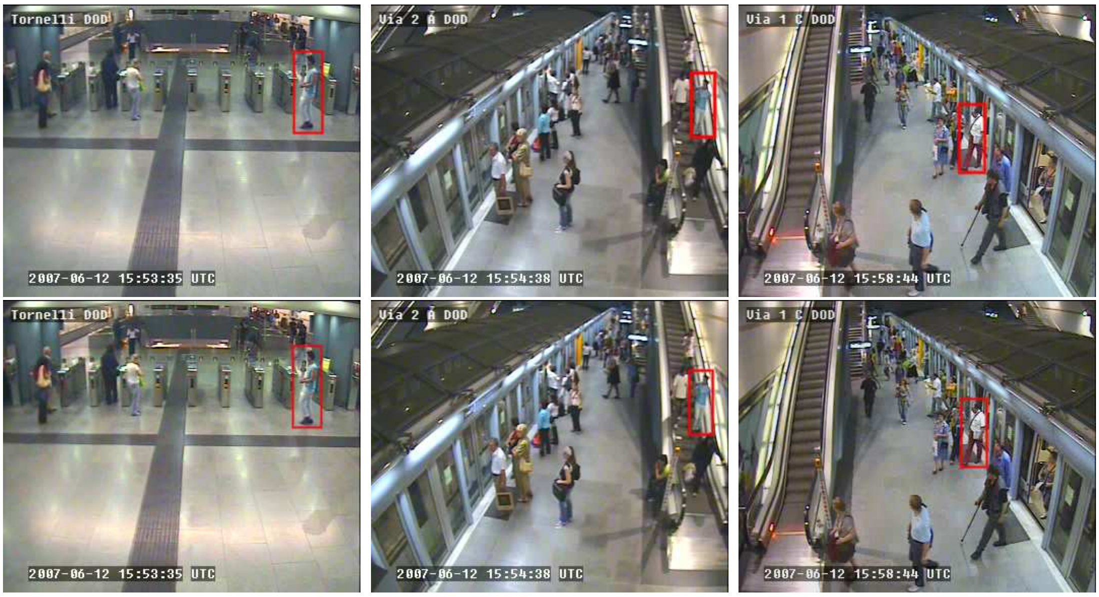


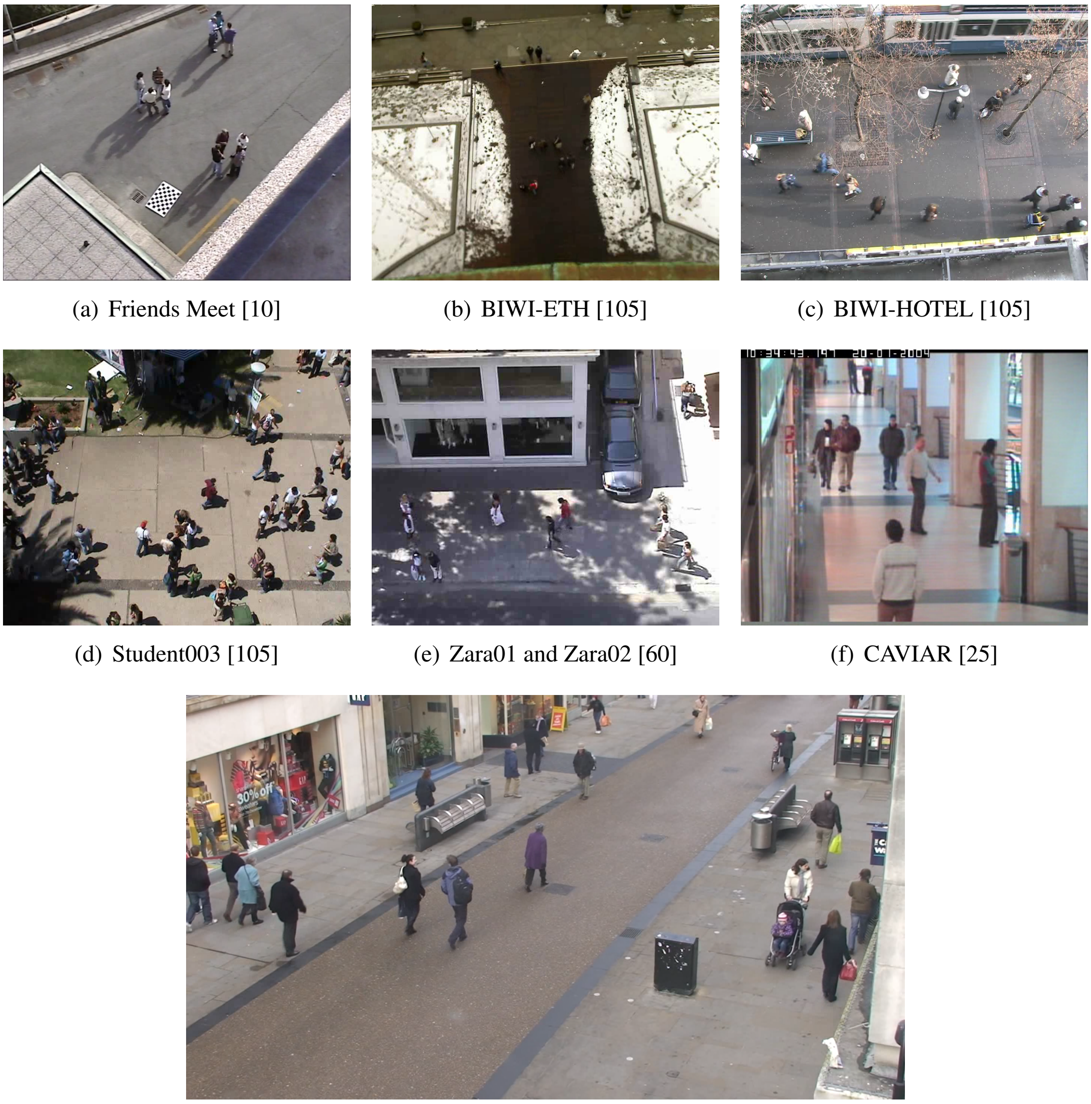
Figure 2.7: Examples of video-based dataset Colombo *et al.* [23] used to test re-identification methods.

Centre [11] (Fig. 2.8(g)), from a street in Oxford (UK), and CAVIAR datasets are used in Qin and Shelton [86] (Fig. 2.8(f)). These datasets present a low to mid-density crowd with people mainly moving in opposite directions. Finally, Ge *et al.* [30, 31] propose SU1, SU2, ARTFEST, STADIUM1, and STADIUM2 datasets that consist of mid-density crowds moving in indoor and outdoor environments (Fig. 2.9) where the camera positioning varies from mid to high.

## 2.6 Discussion

Human motion models were developed in order to simulate realistic people's movements for crowd analysis, but the recent advances of these models allowed them to be applied to tracking and interaction analysis. Compared to using linear motion models, these applications benefit from the fact that the human models can more realistically describe people's movements also because human interactions with other individuals and the environment are considered. The main limitations of these human motion models are the increase in the overall complexity of the application and the need for a parameter tuning phase in order to adapt the specific model to the specific scenario that is being analysed.

Recently, re-identification has been tackled using many different approaches, but it still provides open problems because of the challenges related to the variety of exiting camera networks. In particular, large camera networks, cameras with different specifics, pose changes and illumi-



(a) Friends Meet [10]

(b) BIWI-ETH [105]

(c) BIWI-HOTEL [105]

(d) Student003 [105]

(e) Zara01 and Zara02 [60]

(f) CAVIAR [25]

(g) TownCentre [11]

Figure 2.8: Examples of different datasets used to test group detection and tracking methods.

nation conditions, different positioning of the cameras, and different scenes make the problem difficult to be solved in general. When analysing re-identification algorithms, results on methods solely based on appearance usually achieve less than 40-50% [121] for the first ranking position (the real re-identification score) when 476 images of 119 people are considered in the i-LIDS-static dataset. Re-identification algorithms that exploit also spatio-temporal features can achieve better results for the first ranking position in scenarios where a person's full-body is visible and transitions in the unobserved regions can be linearly modelled. For instance, a re-identification score of over 90% is reported in Javed *et al.* [47] on the dataset in Fig. 2.6(c) where 32 transitions were employed for testing. When human motion models are used to propagate people's

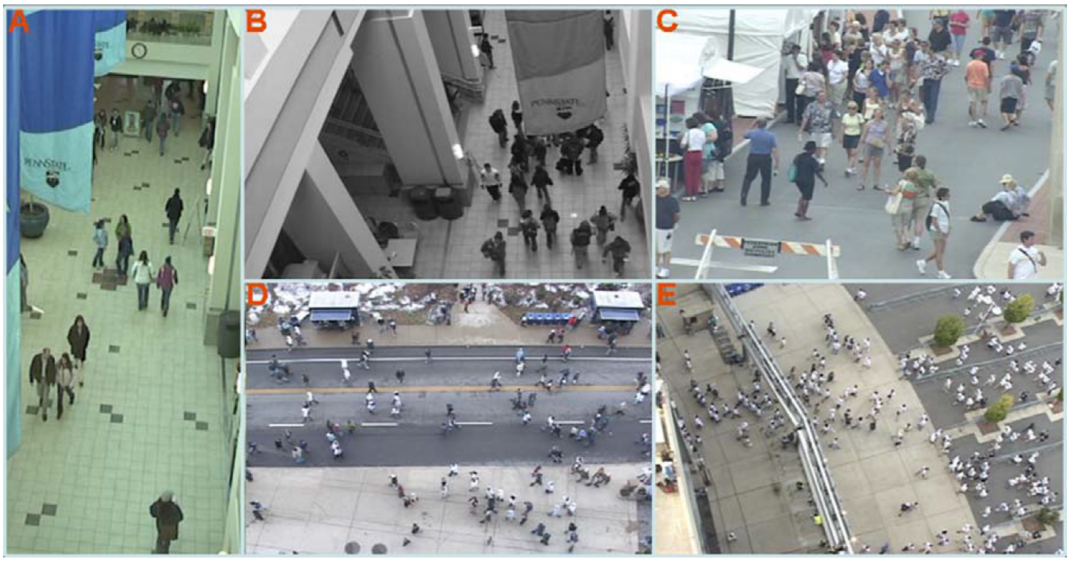


Figure 2.9: Datasets used in Ge *et al.* [30, 31] to test their group detection algorithm. A: SU2, B: SU1, C: ARTFEST, D: STADIUM1, and E: STADIUM2.

movements in the unobserved regions, better performance has been obtained compared to linear propagation and average travelling times [J2]. However, methods solely based on appearance can be tested using single snapshots of people, and they become very important when cameras are located far apart. In fact, spatio-temporal features are less reliable and difficult to use in cases where cameras are located at a considerable distance.

The automatic localisation of groups is very challenging in crowded scenarios, such as public squares or large malls, where spatial proximity alone does not help to determine whether or not people are interacting. Most of the works in the literature use trajectory analysis and clustering [31, 86], however human motion models can be used to analyse people's movement. In particular, compared to methods solely based on people's trajectories, motion models can provide an expected movement for both interacting and non-interacting people, thus providing an understanding of which of these events are taking place within the scene.

# Chapter 3

# Motion prediction

## 3.1 Introduction

Wide sites are extensively monitored by networks of cameras whose field-of-views do not necessarily overlap and, in the presence of unobserved areas, there are no direct measurements available that can be used to facilitate the tracking of a person across cameras. Predicting the exact position where a person exiting the field-of-view of a camera will reappear in the next camera is very challenging due to the presence of various obstacles (barriers and walls) and potential interactions occurring in the unobserved regions. Moreover, in the presence of a crowd or in scenarios where the cameras are at low/mid positioning (Sec. 1.2), partial and complete occlusions will generate challenging situations. Additional challenges are due to changes in illumination conditions across cameras (e.g. the presence of a large window versus an area with artificial illumination only), clutter (different people can look very similar) and different body poses.

In this chapter, we propose an online algorithm that tackles the problem of person re-identification across cameras by exploiting the top-view of the environment representing observed and unobserved regions. In Sec. 3.2, the proposed approach is formalised. We then create a set of possible re-identification candidates for reappearance position and time of each person exiting the field-of-view of one camera by modelling their path in the unobserved regions. In Sec. 3.3.1, the model is based on landmark points (regions of interest or crossing regions) in the scene. In Sec. 3.3.2, the model is based on a goal-driven approach where a set of possible goals are assigned to each person [73, 101], and goals are defined as interest points in the site such as for example shops, doors, exits, seats. In order to propagate people's movement in unobserved areas, we use a motion model developed in the field of crowd simulation [50]. Each person is modelled as an agent that can freely move onto the top-view map trying to reach the selected goals, avoiding barriers and walls while maintaining a desired speed. In Sec. 3.5, association of people across cameras is performed on the possible candidates using spatial and appearance features extracted from the upper body (Sec. 3.4).

Section 3.6 reports the evaluation of the proposed re-identification algorithms based on appearance features only, spatio-temporal features only and a combination of them both. We validate and analyse the results of the goal-driven approach in Sec. 3.6.2, while the proposed person patch is validated in Sec. 3.6.4. In Sec. 3.6.3, we show that the proposed landmark-based method gives the best performances for person re-identification when a combination of appearance and spatial cues are used for association.

## 3.2 Formalisation of the proposed approach

Let $\mathcal{M}$ be a top-view map of the site under surveillance that includes areas observed as well as areas unobserved by the FOVs of $M$ cameras $C_1, C_2, \ldots, C_M$ with non-overlapping FOVs (Sec. 1.3). Observed areas are mapped in $\mathcal{M}$ by homography projection [37]. Let $(x,y) \in \mathcal{M}$ be a point in the top view. Let $N$ people $P_1, P_2, \ldots, P_N$ walk onto $\mathcal{M}$ and let $\mathbf{p}_i(t) = (x_i(t), y_i(t)) \in \mathcal{M}$ be the position of person $P_i$ at time $t$. Finally, let $B \in \mathcal{M}$ be the set of points $\mathbf{p}_B = (x_B, y_B)$ corresponding to barriers and walls that people cannot cross.

We indicate with $\mathbf{p}_i^h(t) = (x_i^h(t), y_i^h(t)) \in \mathcal{M}$ the position of person $P_i$ within the FOV of camera $C_h$, where $t \in [T_{s_i^h}, T_{e_i^h}]$ is the time interval during which $P_i$ is visible in $C_h$. Without loss of generality, we consider camera $C_1$ to be the first camera where the person appears in the scene (i.e. we know $\mathbf{p}_i^1(t)$ with $t \in [T_{s_i^1}, T_{e_i^1}]$) and $C_2$ the second camera where the person reappears. When $t > T_{e_i^1}$, we assume that $P_i$ is not in the FOV of any camera and we start estimating the movement of $P_i$ with the aim of modelling the possible paths to go from $\mathbf{p}_i^1(T_{e_i^1})$ to $\mathbf{p}_i^2(T_{s_i^2})$. Note that $\mathbf{p}_i^1(T_{e_i^1})$ and $\mathbf{p}_i^2(T_{s_i^2})$ may vary for each person. People are expected to move towards regions of interest in the site, a reasonable assumption in those scenarios where areas that most people traverse or reach are straightforward to identify (e.g. exits, seats, lifts in an airport). Moreover, since our focus is to perform motion prediction, we deal with situations where each person that exits camera $C_1$ reappears in camera $C_2$.

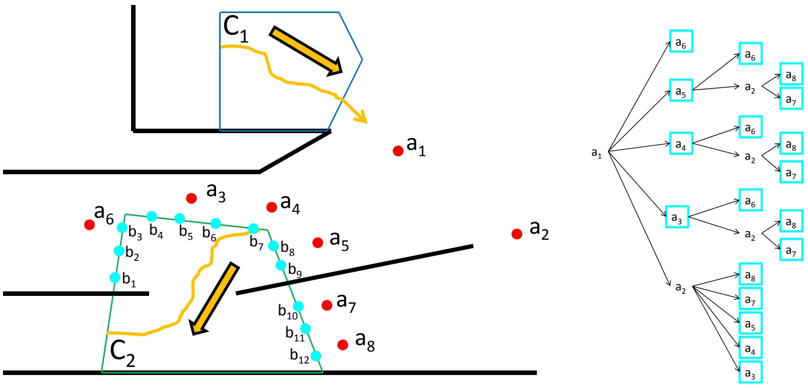
## 3.3 Motion models

### 3.3.1 Landmark-based model

Since in challenging scenes the travelling time and entry/exit regions of the FOVs are not sufficient for person re-identification, we propose to create person re-identification candidates by modelling potential people's movements in unobserved regions using areas of interest and crossing areas (landmarks) obtained using an environment map of the site under surveillance and, in our implementation, manually defined on the top view [J3]. We define *crossing landmarks* as the regions through which people transit and *entry landmarks* as the regions where people may enter the FOV of the next camera (Fig. 3.1(a)). We refer to the proposed motion propagation based on regions of interest as the Landmark-Based Model (LBM).

When person $P_i$ exits $C_1$, its movement is propagated towards a first landmark, and then towards crossing and entry landmarks according to specific transition rules. These rules define how people can move through the crossing landmarks and which entry landmarks can be reached by a crossing landmark (Fig. 3.1(b)). The entry landmarks reached after the transitions are the candidate areas for reappearance of $P_i$ in $C_2$. Notice that landmarks are fixed for a specific map and, in absence of other prior information, we assume that each landmark is equally likely to be traversed in the propagation. A time step is associated to each reached entry landmark, and calculated by the speed equation using the speed of people registered in the first observed region and the distance covered by the propagation through the landmarks on the top view $\mathcal{M}$. With this modelling, variations of people's speed occurring in the unobserved regions are not explicitly taken into account, however experimental results (Sec. 3.6.3) demonstrate that the model can also cope with changes in people's speed, to some extent. Figure 3.2 shows the block diagram of the proposed method.

We model crossing and entry landmarks with a set of vertices $V$ of an oriented graph $G = (V,E)$, where $E$ is the set of oriented edges that connect the vertices and correspond to the transitions across landmarks in $\mathcal{M}$ (Fig. 3.1(b)). Let $l(\iota)$, with $\iota \in E$, be the length of $\iota$. Let $A_V = \{a_1, a_2, \ldots, a_{|A_V|}\}$ with $A_V \subseteq V$ and $|A_V| > 0$, be the set of crossing landmarks and $B_V = \{b_1, b_2, \ldots, b_{|B_V|}\}$ with $B_V \subseteq V$ and $|B_V| > 0$, be the set of entry landmarks. Let $F_V = \{f_1, f_2, \ldots, f_{|F_V|}\}$ with $F_V \subseteq A_V$ and $|F_V| > 0$, be the set of vertices where the propagation of



(a) (b)

Figure 3.1: Landmark-Based Model (LBM): (a) Example of setup (black line: environment map; blue and green line: FOVs of Camera 1 ($C_1$) and Camera 3 ($C_2$) at London Gatwick airport, respectively; orange line: trajectory; orange arrow: direction of motion; red dot: crossing landmark; cyan dot: entry landmark). (b) Transition graph ($\{a_1, a_2, \ldots, a_8\}$: crossing landmarks; cyan border: crossing landmark from where entry landmarks can be reached); $a_1$ and $a_2$ do not have a cyan border as entry landmarks cannot be reached from them.

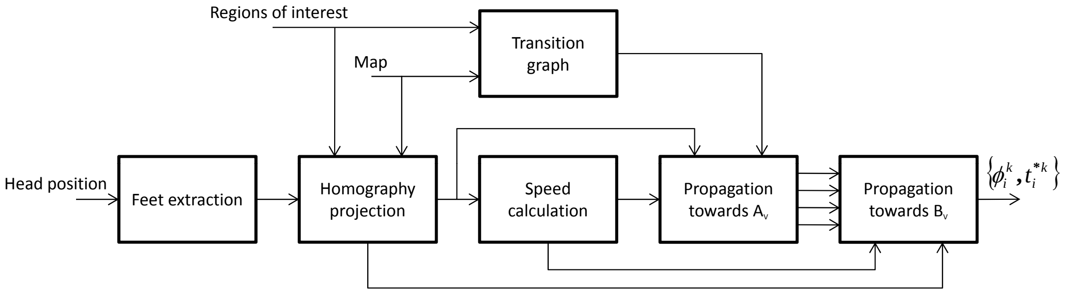


Figure 3.2: Block diagram for the proposed Landmark-Based Model (LBM).

people's movements can start from. Let us define

$$E_i^* = E \cup \{(\mathbf{p}_i^1(T_{e_i^1}), F_V)\}, \tag{3.1}$$

where $(\mathbf{p}_i^1(T_{e_i^1}), F_V)$ corresponds to the edges connecting $\mathbf{p}_i^1(T_{e_i^1})$ to the set of vertices in $F_V$, namely the connection between the last visible position of $P_i$ in $C_1$ and the vertices where the propagation can start from. Then, let us define the ordered set of edges that a person can follow to go from $\mathbf{p}_i^1(T_{e_i^1})$ to all the entry landmarks $v \in B_V$ as

$$\phi_i^k = \left(\iota_1, \iota_2, \ldots, \iota_h, \ldots, \iota_{|\phi_i^k|}\right), \tag{3.2}$$

where $\iota_h \in E_i^*$; $k = 1, 2, \ldots, |\Phi_i|$ and $\Phi_i = \{\phi_i^k\}$ is the set of all possible paths that person $P_i$

can follow; $\iota_1 = (\mathbf{p}_i^1(T_{e_i^1}), F_V)$ is the first edge of the sequence; $\iota_{|\phi_i^k|} = (A_V, B_V)$ indicates that the last edge of $\phi_i^k$ must go towards an entry landmark; and the edges from $\iota_2$ to $\iota_{|\phi_i^k|-1}$ are selected according to the transition rules. We now accumulate the time needed for person $P_i$ to travel through a possible path $\phi_i^k$ using the speed equation

$$t_{\phi_i^k} = \sum_{h=1}^{|\phi_i^k|} \frac{l(\iota_h)}{s_i}, \tag{3.3}$$

where $s_i$ is the maximum speed calculated within a time window of $T_p$ frames in $C_1$. The sum of the time step when person $P_i$ exits $C_1$, $T_{e_i^1}$, and the time required for $P_i$ to traverse $\phi_i^k$, $t_{\phi_i^k}$, defines the time step when person $P_i$ reaches $C_2$ if $\phi_i^k$ is traversed

$$t_i^{*k} = T_{e_i^1} + t_{\phi_i^k}. \tag{3.4}$$

The above process is repeated for each person exiting $C_1$ and going to $C_2$. When person $P_r$, with $r = 1, 2, \ldots, N$, reappears in $C_2$, the set of candidates for the association are the set of vertices $V_i^* = \{v_i^{*k}\} \in B_V$ reached by $\iota_{|\phi_i^k|}$ that satisfy

$$T_{s_r^2} - \Delta_t < t_i^{*k} < T_{s_r^2} + \Delta_t, \tag{3.5}$$

where $\Delta_t \in \mathbb{N}$, thus restricting the set of possible candidates from person $P_i$ to the closest in time to the reappearance of $P_r$. If $\Delta_t$ is too small, the time window would be too restrictive and the method could not account for small variations in speed. If $\Delta_t$ is too large, the time window would lose its significance to select only the "good" candidates for re-identification.

### 3.3.2 Social Force Model-based model

*Overview*

In this section, we propose to exploit the SFM to create person re-identification candidates. SFM is initialised for each person with information from the first observed region and then the path of a person is propagated within the unobserved areas on $\mathcal{M}$. Since in a complex site people have different goals to reach, a unique fixed goal for all the people is not a good model for the estimation of their behaviour [101]. Unlike Luber *et al.* [67] that employs an extensive training to understand where people are likely to move, we tackle this problem by introducing a Multi-Goal Social Force Model (MG-SFM) [J2] that spreads $|G|$ different goals corresponding to interest

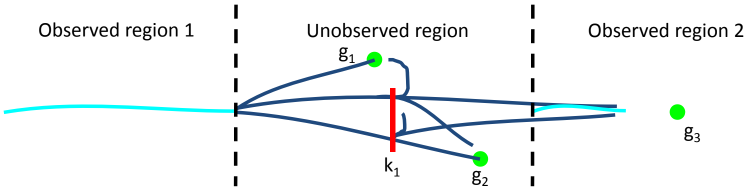


Figure 3.3: Schematic representation of the evolution of a path using the proposed MG-SFM. Cyan line: trajectory in the observed region. Green dot: goal. Red line: key region where new predictions are generated towards the goals. Blue line: predicted trajectory towards the goal.

points in the site, such as shops, cafeterias, exits, seats, etc.

Let us define $\Psi_i^*(t) = \{\mathbf{p}_i^{*j}(t)\}$ where $\mathbf{p}_i^{*j}(t) \in \mathcal{M}$, $j = 1, 2, \ldots, |\Psi_i^*(t)|$, and $|\Psi_i^*(t)|$ is the number of position candidates at time $t$ where the person $P_i$ is likely to walk. For each $\mathbf{p}_i^{*j}(t) \in \Psi_i^*(t)$ a goal $\mathbf{g}_i^j$ is fixed where $\mathbf{g}_i^j \in G$ and $G$ is the set of possible goal positions (interest points) onto $\mathcal{M}$ (in our implementation, goal positions are manually defined). $\mathbf{p}_i^{*j}(t)$ and $\mathbf{g}_i^j$ will be considered as a pair in the rest of the thesis. As it is difficult to exactly define the desired goal of each person over time, we generate candidates of people's movement by introducing a set of new predictions towards $G$ when the already existing trajectories in $\Psi_i^*(t)$ reach key regions in the environment (i.e. a crossing of possible paths selected using the map of the environment), represented by $\mathcal{K} = \{\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_{|\mathcal{K}|}\}$ where $|\mathcal{K}|$ is the number of key regions in $\mathcal{M}$. Note that due to the different way they are modelled, these key regions may not be the same as the landmarks defined in Sec. 3.3.1.

Without loss of generality, we assume $C_2$ to be the next camera where person $P_r$, with $r = 1, 2, \ldots, N$, is visible ($T_{s_r^2}$ is the time step when $P_r$ reappears). We consider all the predictions $\mathbf{p}_i^{*j}(t)$ at time $t \in [T_{s_r^2} - \Delta_t, T_{s_r^2} + \Delta_t]$, where $\Delta_t$ is a time interval, and we set their next goal to $\mathbf{p}_r^2(T_{s_r^2})$. $\Delta_t$ is the same as defined in Sec. 3.3.1. Then, we let the predictions evolve over time along with the new observed trajectory (the new goal) $\mathbf{p}_r^2$ for $T_\pi$ frames. Finally, from all $\mathbf{p}_i^{*j}$ we select the closest prediction in space to $\mathbf{p}_r^2$ in order to re-identify $P_r$ (ideally $P_r$ is re-identified with $P_i$ when they represent the same person).

Figure 3.3 shows examples of predictions obtainable with the proposed approach: the algorithm finds the next position of a pedestrian starting from the observations in the first camera and uses this information to estimate the path a person is expected to follow when observations are available again in the next camera.

### *Multi-Goal Social Force Model*

We modify the Social Force Model by modelling people's movement towards specific goals, avoiding barriers and walls, and maintaining a desired speed that we calculate in the first observed region. We assume that there are no significant interactions between people and, as for LBM, we do not explicitly model variations of people' speed when they cross the unobserved regions. Let each person $P_i$ have mass $m_i$ and be guided by the forces that describe the desired movements according to the surrounding constraints. We model an attractive force $\mathbf{f}^{*j}_{D_i}(t)$ towards a specific goal and a repulsive force $\mathbf{f}^{*j}_{B_i}(t)$ as the sum of forces from walls and barriers. Finally, the displacement of $P_i$ over time is defined by $\frac{d\mathbf{v}^{*j}_i(t)}{dt}$. The dynamic of the SFM is therefore formulated as

$$m_i \frac{d\mathbf{v}^{*j}_i(t)}{dt} = \mathbf{f}^{*j}_{D_i}(t) + \mathbf{f}^{*j}_{B_i}(t). \tag{3.6}$$

As abrupt movements of walking people are less likely to happen, we define a temporal smoothing process similar to the one reported in [93] in order to estimate the next step by considering the velocity[1] in the previous steps and actual forces. Compared to [93], we use a weighted average of the two components and we use more than only one previous step for smoothness

$$\mathbf{p}^{*j}_i(t+1) = \mathbf{p}^{*j}_i(t) + \left( w \frac{d\mathbf{v}^{*j}_i(t)}{dt} \tau + (1-w)\overline{\mathbf{v}}^{*j}_i(t) \right), \tag{3.7}$$

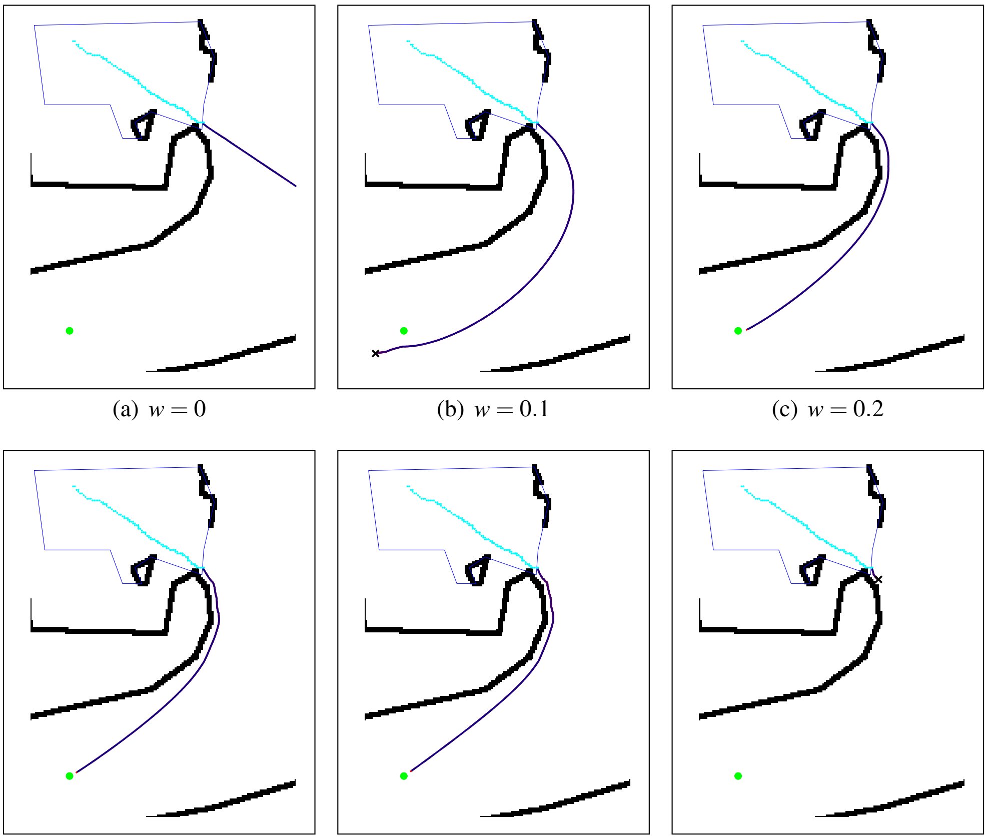
where $\overline{\mathbf{v}}^{*j}_i(t) = \frac{\mathbf{p}^{*j}_i(t) - \mathbf{p}^{*j}_i(t-T_p)}{T_p}$ is the actual velocity calculated as the average velocity of the previous $T_p$ frames, $\tau$ is the interval during while the variation of velocity is calculated. The magnitude of the displacement is directly proportional to $\tau$. We fix $\tau = 1$ as we calculate $\tau$ at each time step. $w \in [0,1]$ is the weight given to the actual velocity and $1-w$ the one given to the previous velocity. The movement smoothness is inversely proportional to $w$ and high values of $w$ can lead to abrupt displacement of the target over time. Figure 3.4 shows different trajectory behaviours at varying $w$.

A goal is a point or an area of interest that would be reached at a desired speed following the minimum path, if there would not be any constrain such as walls and barriers. These desires are taken into account as

$$\mathbf{f}^{*j}_{D_i}(t) = m_i \frac{v^0_i \mathbf{e}^{0*j}_i(t) - \overline{\mathbf{v}}^{*j}_i(t)}{\tau_i}, \tag{3.8}$$

where $v^0_i$ is the desired speed towards the direction $\mathbf{e}^{0*j}_i(t)$ of the goal to reach, and $\tau_i$ is the time

[1] Velocity is the 2D displacement of a point, while speed is the magnitude of the velocity.



(a) $w = 0$ (b) $w = 0.1$ (c) $w = 0.2$

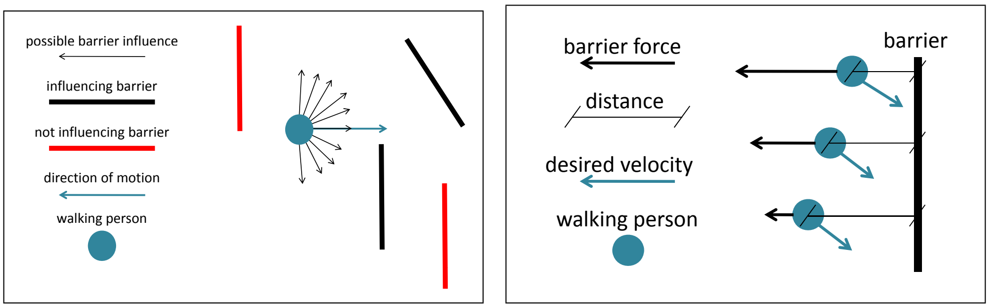
(d) $w = 0.3$ (e) $w = 0.4$ (f) $w = 0.5$

Figure 3.4: Trajectory propagation examples generated using different values of $w$ (see Eq. 3.7 for details) on the top-view map. Black line: barrier. Cyan line: trajectory from the observed region (FOV of the camera is the blue region). Purple line: propagated trajectory. Green dot: goal to reach. Black cross: example of stopped prediction because its speed is too slow.

relaxation parameter. $\mathbf{f}^{*j}_{D_i}(t)$ is the force that pushes the target to reach the desired velocity by calculating the difference between desired and actual velocities. Note that $v_i^0$ does not depend on the specific $j$ prediction but only on the desired speed of person $P_i$.

The desired speed $v_i^0$ is a key feature for our model. We have tested three different strategies for desired speed calculation using observations from the first observed region: the average speed using the complete trajectory here referred to as MG-SFM-AVG; the maximum speed registered within a time interval of $2 \cdot T_p$ (MG-SFM-MAX50); the maximum speed registered within a time interval of $T_p$ (MG-SFM-MAX25), called $s_i$ in Sec. 3.3.1. Results for the three strategies are reported in Fig. 3.11 and explained in Sec. 3.6.2.

A monotonically decreasing force $\mathbf{f}^{*j}_{B_i}(t)$ is also considered that acts from barriers and walls to that person [50]. As suggested in [50], we model this force with an inverted exponential proportional to the Euclidean distance $d^{*j}_{B_i}(t)$ between person $P_i$ predictions and barriers $B$. In



(a) (b)

Figure 3.5: Influence of the presence of barriers on the movement of a person: (a) the influence is limited to the visible range $[-90°, 90°]$ in the direction of motion; (b) the force generated by a barrier is inversely proportional to the exponential of the distance from the barrier itself [50].

addition to this, as walking people are influenced only from what happens in front of them [50], we restrict $\mathbf{f}^{*j}_{B_i}(t)$ to the barriers in the range $[-90°, 90°]$ of the direction of motion of $P_i$ and to the "visible" barriers from the actual position of the pedestrian. Figure 3.5(a) shows the range of influence of the barriers on a person and Fig. 3.5(b) reports a schematic representation of the influence of barrier forces on people's movement, formalised as

$$\mathbf{f}^{*j}_{B_i}(t) = \sum_B A_B e^{-\frac{\mathbf{d}^{*j}_{B_i}(t)}{B_B}}, \tag{3.9}$$

where $A_B$ is the weight associated to the barrier force (high values correspond to high repulsion force from the barriers), $B_B$ is the interaction range that enlarges or reduces the area of influence of the barriers on people's movements, $\mathbf{d}^{*j}_{B_i}(t)$ is a force where the magnitude is the Euclidean distance value and the direction is the direction between propagation, $j$, and barrier, and the summation indicates the sum of forces from each barrier position.

We predict how each person moves towards each goal using Eq. 3.7. At time step $T_{e^1_i} + 1$ (when person $P_i$ is no longer visible from camera $C_1$), we generate $|G|$ predictions towards each goal in $G$ and we let them propagate onto $\mathcal{M}$. Since walking people change their view of the environment, it is likely that the direction of motion towards their goal changes over time. To model this behaviour, multiple new predictions are further generated when an existing prediction reaches any key region $\mathbf{k}$. For instance, if prediction $\mathbf{p}^{*1}_1(\bar{t})$ towards goal $\mathbf{g}^1_1$ has reached the key region $\mathbf{k}_1$ at time $\bar{t}$, we generate $|G|-1$ new predictions towards $G/\left\{\mathbf{g}^1_1\right\}$ (we exclude the goal already followed by $\mathbf{p}^{*1}_1(\bar{t})$), and we include[2] them in $\Psi^*_1(\bar{t})$. The next step of MG-SFM removes

[2] For the new predictions, we include in $\Psi^*_1(t)$ the same positions of $\mathbf{p}^{*1}_1(t)$ for $t = [T_{e^1_1}+1, \bar{t}]$, and from

Table 3.1: Testing for MG-SFM predictions on 42 trajectories from the London Gatwick airport dataset. See text for the complete explanation of MG-SFM-AVG, MG-SFM-MAX50, and MG-SFM-MAX25. Columns 2-4: Average percentage of predictions within the indicated radius centred on the position of person's reappearance of the 60 closest predictions (in time) to the reappearance time step. Columns 5-7: Average percentage of time synchronisation within the indicated frames between predictions and time step of person's reappearance of the 60 closest predictions to the position of reappearance.

| | **Radius (units)** | | | **Time (frames)** | | |
|---|---|---|---|---|---|---|
| | **5** | **10** | **20** | **25** | **75** | **125** |
| **MG-SFM-AVG** | 45% | 57% | 67% | 31% | 62% | 71% |
| **MG-SFM-MAX50** | 48% | 59% | 71% | 33% | 79% | 86% |
| **MG-SFM-MAX25** | 45% | 62% | 81% | 52% | 79% | 90% |

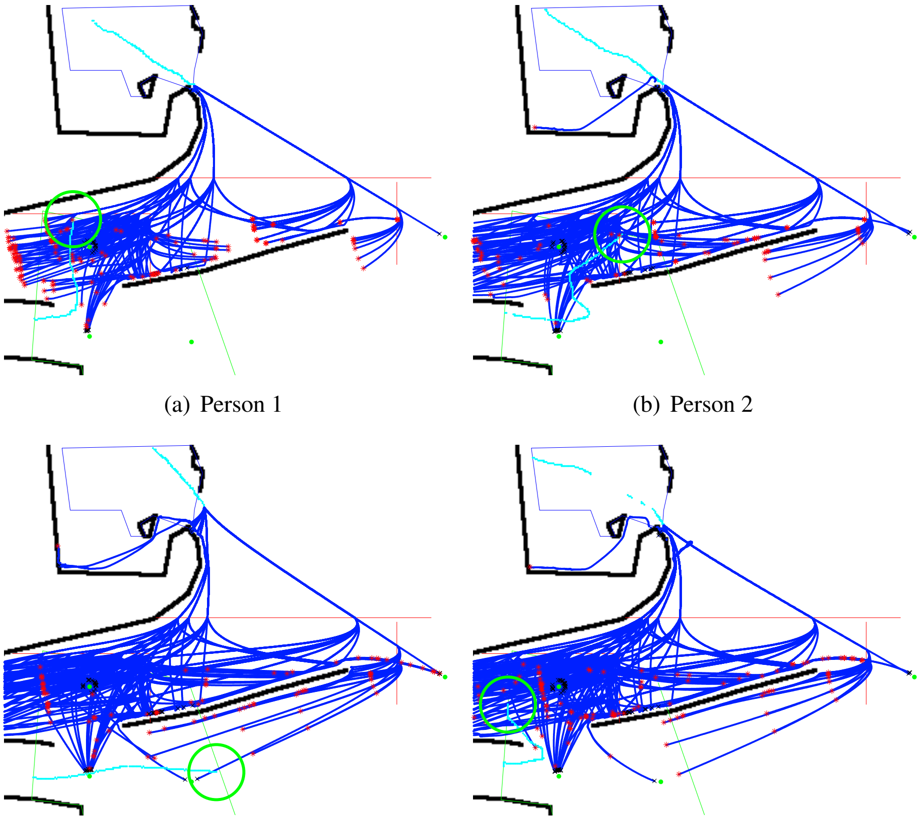
from $\Psi_i^*(t)$ the predictions that do not appropriately model realistic scenarios. In particular, we remove each $\mathbf{p}_i^{*j}(t)$ with distance from its goal $\mathbf{g}_i^j$ less than $\varepsilon_g > 0$, and we remove each $\mathbf{p}_i^{*j}(t)$ that corresponds to a prediction with speed $v_i^{*j}(t) < \varepsilon_v \cdot v_i^0$, where $v_i^{*j}(t) = |\overline{\mathbf{v}}_i^{*j}(t)|$ and $0 < \varepsilon_v < 1$.

Figure 3.6 shows four examples of trajectory prediction in unobserved regions using the parameter setting explained in Sec. 3.6.1. Using the same parameter setting, we test our model in order to calculate the distance (in time and space) of our predictions with respect to frame step and position of a person's reappearance. Table 3.1 shows the results for MG-SFM-AVG, MG-SFM-MAX50, and MG-SFM-MAX25 calculated on 42 people going from one observed region to the next. For each person we consider the 60 closest predictions in time to the reappearance time step and we calculate the average distance to the reappearance position. The results are shown in columns 2-4 of Table 3.1. We see that for MG-SFM-MAX25, 81% of the predictions are within 20 units (as the radius of the green circle in Fig. 3.6). Furthermore, we analyse how synchronised our predictions are with the reappearance time step. We take the 60 closest predictions in space to the position of reappearance and we calculate the average difference with the time step of reappearance. Columns 5-7 of Table 3.1 summarise the results, where we can see that over 50% of our predictions are within 25 frames (1 second on the used dataset) when applying MG-SFM-MAX25.

As predicting the exact position and the exact time instant when a person reappears is very challenging, when a generic person $P_r$ appears in $C_2$ we consider good candidates for $P_r$ all the predictions

$$\mathbf{p}_i^{*j}(t) \in \Psi_i^*(t) \tag{3.10}$$

$\bar{t}+1$ onward we make the predictions towards the assigned goals.



(a) Person 1

(b) Person 2

(c) Person 3

(d) Person 4

Figure 3.6: Examples of trajectory prediction for four people walking from Camera 1 ($C_1$) to Camera 3 ($C_2$) of the i-LIDS dataset from London Gatwick airport [43]. Cyan line: trajectory in the observed regions. Blue line: predicted trajectory using MG-SFM-MAX25 (see text for details). Red star: predicted trajectory at the time step when the person reappears in $C_2$. Black cross: predicted trajectory that stops because it reached the goal or its speed is too small. Red segment: definition of the key region for splitting the predictions. Black segment: barrier. Green dot: goal. Green circle: 20 units of radius centred in the first observation in $C_2$.

from $P_i$ where $i = 1,2,\ldots,N$, $j = 1,2,\ldots,|\Psi_i^*(t)|$, $t \in [T_{s_r^2} - \Delta_t, T_{s_r^2} + \Delta_t]$ and $\Delta_t = 2 \cdot T_p$. $\Delta_t$ is chosen to be proportional to $T_p$ (Eq. 3.7) in order to obtain a large enough time window for the final association between predictions $\mathbf{p}_i^{*j}(t)$ and observations $\mathbf{p}_r^2(t)$ (similarly to Sec. 3.3.1). Note that $P_i$ varies among all the available people's trajectories within the specific time interval since we are now tackling the association problem. In particular, we propagate the predictions $\mathbf{p}_i^{*j}(t + t_r)$ towards $\mathbf{p}_r^2(T_{s_r^2} + t_r)$ with $t_r = 0, 1, \ldots, T_\pi - 1$ using Eq. 3.7, where

$$T_\pi = \min(T_{e_r^2} - T_{s_r^2} + 1, T_p). \tag{3.11}$$

In other words, good candidate predictions for person $P_r$ consider $\mathbf{p}_r^2(t)$ as their goal for $T_\pi$ frames. Algorithm 1 reports the complete algorithm for the MG-SFM.

**Algorithm 1** MG-SFM for camera pairs

Define: map $\mathcal{M}$; set $B$ of barrier positions; goals $\mathbf{g} \in G$; set of key regions $\mathcal{K}$; parameters $\varepsilon_v$ and $\varepsilon_g$; $T$: set of considered time steps; $I$: set of walking people; $T_p$: frames to consider for actual velocity; $\Delta_t$: frame interval for re-identification;
$C_1$: first observed region; $[T_{s_i^1}, T_{e_i^1}]$: time interval when person $P_i$ is within the FOV of $C_1$;
$\mathbf{p}_i^1(t)$: position of person $P_i$ at time $t$ within $C_1$; $v_i^0$: desired speed of person $P_i$;
$C_2$: second observed region; $[T_{s_i^2}, T_{e_i^2}]$: time interval when person $P_i$ is within the FOV of $C_2$;
$\mathbf{p}_i^2(t)$: position of person $P_i$ at time $t$ within $C_2$;
$\mathbf{p}_i^{*j}(t)$: predicted position of person $P_i$ towards goal $\mathbf{g}_i^j$ at time $t$, $\mathbf{g}_i^j \in G$;
$\mathrm{d}\Big(\mathbf{a}, \mathbf{b}\Big)$: Euclidean distance between $\mathbf{a}$ and $\mathbf{b}$;
$\min\Big(a, b\Big)$: minimum value between $a$ and $b$;

```
for all $t \in T$ do
    for all $i|P_i \in I$ do
        if $t \in [T_{s_i^1}, T_{e_i^1}]$ then                                    ▷ First observed region
            obtain $\mathbf{p}_i^1(t)$ by single-camera tracking
        else                                                                   ▷ Unobserved regions
            if $t = T_{e_i^1} + 1$ then                                         ▷ Initialisation of $\Psi_i^*(t)$
                initialise $\Psi_i^*(t) = \{\mathbf{p}_i^1(t)\}$
                $\Psi_i^*(t) =$ADDBRANCHES$\Big(\Psi_i^*(t), G\Big)$
            end if
            for all $j\,|\,\mathbf{p}_i^{*j}(t) \in \Psi_i^*(t)$ do                        ▷ Prediction step
                apply Eq. 3.7 to $\mathbf{p}_i^{*j}(t)$ (towards $\mathbf{g}_i^j$)
                $v_i^{*j}(t)$ = speed of $\mathbf{p}_i^{*j}(t)$
                if $\Big(t > T_{e_i^2} + T_p \wedge v_i^{*j}(t) < \varepsilon_v \cdot v_i^0\Big) \vee \Big(d\Big(\mathbf{p}_i^{*j}(t), \mathbf{g}_i^j\Big) < \varepsilon_g\Big)$ then      ▷ Check for non-valid predictions
                    $\Psi_i^*(t) = \Psi_i^*(t) / \Big\{\mathbf{p}_i^{*j}(t)\Big\}$
                end if
                if $\mathbf{p}_i^{*j}(t)$ within $\mathcal{K}$ then
                    $\Psi_i^*(t) =$ADDBRANCHES$\Big(\Psi_i^*(t), G/\Big\{\mathbf{g}_i^j\Big\}\Big)$
                end if
            end for
        end if
    end for
    initialise $j_r = 1$, $\overline{\Psi}_i^*(t) = \emptyset$
    for all $r|P_r \in I$ do                                                       ▷ Second observed region
        for all $t^* \in [T_{s_r^2} - \Delta_t, T_{s_r^2} + \Delta_t]\,|\,\exists \mathbf{p}_i^{*j}(t^*) \in \Psi_i^*(t^*)$ do
            $T_\pi$ = $\min\Big(T_{e_r^2} - T_{s_r^2} + 1, T_p\Big)$
            for all $j\,|\,\mathbf{p}_i^{*j}(t^*) \in \Psi_i^*(t^*)$ do
                $\overline{\mathbf{p}}_i^{*j_r}(1 \to t^*) = \mathbf{p}_i^{*j}(1 \to t^*)$
                $\overline{\Psi}_i^*(t^*) = \overline{\Psi}_i^*(t^*) \cup \Big\{\overline{\mathbf{p}}_i^{*j_r}(1 \to t^*)\Big\}$
                for all $t_r \in [0, T_\pi - 1]$ do
                    apply Eq. 3.7 to $\overline{\mathbf{p}}_i^{*j_r}(t^* + t_r)$ (towards $\mathbf{p}_r^2(T_{s_r^2} + t_r)$)
                end for
                $j_r = j_r + 1$
            end for
        end for
    end for
end for

procedure $\Psi$ = ADDBRANCHES($\Psi$,$G$)              ▷ Procedure to add new branches for trajectory prediction
    $\Psi$: set of trajectory predictions; $G$: set of goal positions
    for all $\mathbf{p} \in \Psi$ do
        for all $\mathbf{g} \in G$ do
            create new $\overline{\mathbf{p}} = \mathbf{p}$
            associate $\overline{\mathbf{p}}$ to the goal $\mathbf{g}$
            $\Psi = \Psi \cup \overline{\mathbf{p}}$
        end for
    end for
end procedure
```

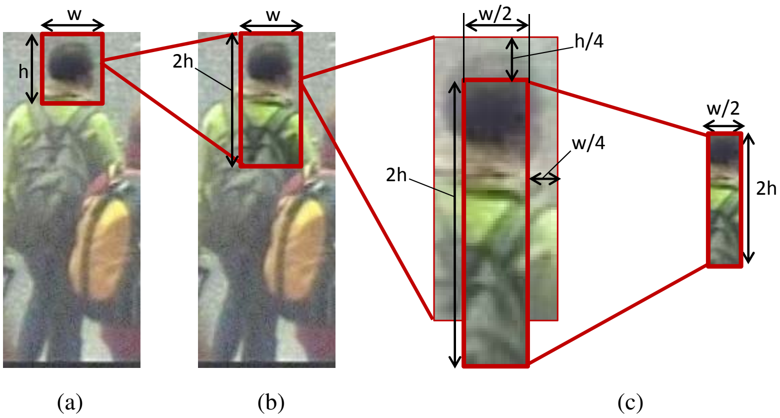


Figure 3.7: Spatial support for person's representation. (a) Head detection bounding box. (b) Selected strip whose height is twice the height, $h$, and width is half the width, $w$, of the bounding box. (c) The strip is shifted downward by $h/4$ to reduce the likelihood of the presence of background pixels in the features used for association. Compared to full-body, this upper-body strip is more effective as a person's representation for re-identification in crowded scenarios (see Sec. 3.6.4 and Fig. 3.15).

## 3.4 Target representation

We introduce a person's representation model for crowded scenarios that is defined as a vertical strip located around the head [J3] as in typical surveillance settings the head and the upper body are the most frequently visible and recognisable part of a person [115]. The top part of the strip is centred on the head and contains pixels of the upper body (Fig. 3.7), thus reducing the probability of occlusion and the presence of the background while maintaining the most discriminative part of each person. People's trajectories on the ground are then created using feet locations that are estimated starting from the head positions, and assuming that people stand upright [68].

The appearance features for our method are extracted from the upper-body strip using a single snapshot for each person (no temporal grouping is performed): we use a concatenation of 16-bin histograms as in [35, 85, 121] (see also Fig. 2.2). In particular, we employ 8 colour channels (R, G, B, Y, Cb, Cr, H, S) from RGB, YCbCr and HSV colour spaces, 8 Gabor filters (with the following parameters: $(\gamma,\theta,\lambda,\sigma^2) = (0.3,0,4,2)$, $(0.3,0,8,2)$, $(0.4,0,4,1)$, $(0.3,\pi/2,4,2)$, $(0.3,\pi/2,8,1)$, $(0.3,\pi/2,8,2)$, $(0.4,\pi/2,4,1)$, $(0.4,\pi/2,8,2)$), and 13 Schmid filters (with the following parameters: $(\sigma,\tau) = (2,1)$, $(4,1)$, $(4,2)$, $(6,1)$, $(6,2)$, $(6,3)$, $(8,1)$, $(8,2)$, $(8,3)$, $(10,1)$, $(10,2)$, $(10,3)$, $(10,4)$).

## 3.5 Association

When a person $P_r$ becomes visible in $C_2$, the association between $P_r$ and the re-identification candidates generated by people $P_i$ that exited $C_1$, has to be performed. We propose to use a reappearance score given by a weighted sum of the ranking obtained by measuring the spatial Euclidean distance between candidates and reappearance positions, $\mathbf{p}_r^2(T_{s_r^2})$, on $\mathcal{M}$, and the ranking of the appearance similarity measures.

The Euclidean distance is calculated, in the case of LBM (Sec. 3.3.1), between candidates $v_i^{*k}$ and $\mathbf{p}_r^2(T_{s_r^2})$, while in the case of MG-SFM (Sec. 3.3.2), we define $d_{ir}^{*j}(t)$ to be the Euclidean distance between $\mathbf{p}_i^{*j}(t+T_\pi)$ and $\mathbf{p}_r^2(T_{s_r^2}+T_\pi)$, where $\mathbf{p}_i^{*j}(t+T_\pi)$ is defined in Eq. 3.10 and $T_\pi$ in Eq. 3.11, and for each $P_i$ we calculate

$$\chi_{ir} = \min_j \min_t \left( d_{ir}^{*j}(t) \right), \tag{3.12}$$

in order to consider only the best reappearance candidate for each $P_i$. The ranking based on spatio-temporal cues of the re-identification candidates for $P_r$ is calculated by sorting the Euclidean distances of the various candidates. Since the association using appearance information is performed separately, we can use any appearance-based method. From the state-of-the-art methods [28, 84, 85, 121], we choose Bhattacharyya distance as an association measure because it does not require any learning phase, and it outperforms both L1-Norm and rankSVM distances when applied to colour and texture features (Sec. 3.4). Also in this case, the ranking based on appearance cues is calculated by sorting the specific distance applied.

## 3.6 Results and analysis

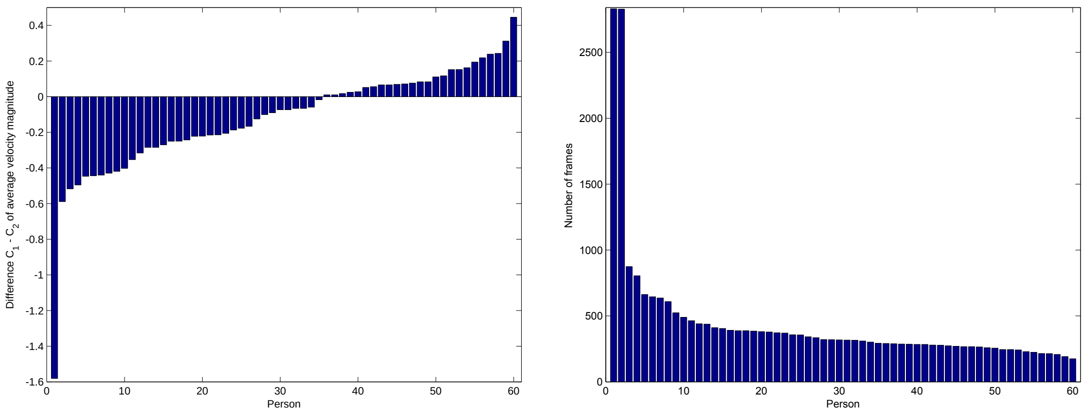
### 3.6.1 Experimental setup

To validate the proposed methods, we use the i-LIDS dataset from London Gatwick airport [43] and we study the movement of people at the arrival terminal. We consider people that are visible when they walk out of the *passengers area*. The aim is to find where and when these people reappear in one of the next cameras in the *public area*. This is a challenging environment where people can potentially walk in many directions once they exit the camera view covering the passenger area, and movements may be constrained by barriers. In addition, cameras present large illumination changes and people can reappear with different poses after transiting in the unobserved regions where different paths can be followed. In the experiments, Camera 1 is the first observed region ($C_1$) and Camera 3 of the dataset is the second camera where people reappear ($C_2$). We only use Camera 1 and Camera 3 because people's locations for Camera 2, Camera 4 and Cameras 5 are not available. The top-view map $\mathcal{M}$ is shown in Fig. 1.1[3] and we consider 60 people similarly to previous works [47]. Results are shown by Cumulative Matching Characteristic (CMC) curve where the ideal result is a horizontal line at value 1 that corresponds to having correct re-identification for all people[4]. We use $\Delta_t = 50$ frames (corresponding to two seconds) for association. The performance of the motion predictions does not substantially change their performance by varying $\Delta_t$: for LBM, for instance, we obtain a mean re-identification score of 31% and standard deviation of 3% by varying $\Delta_t$ from 20 to 80 frames at steps of 5 frames.

To better understand the variation in people's movement and the travelling time variability between $C_1$ and $C_2$, Fig. 3.8 and Fig. 3.9 show some statistics obtained using ground-truth information on the top view from 60 people. Figure 3.8(a) reports the difference of the average speed (velocity magnitude) registered in $C_1$ and $C_2$, showing how people move at substantially different speeds. Figure 3.8(b) shows the travelling time to go from $C_1$ to $C_2$, and Fig. 3.9 shows the colour-coded time evolution of people in the two cameras where segments correspond to time intervals during which a person is in the FOV of a camera and is not totally occluded. It is interesting to note that *(i)* some people stay in the FOV of $C_1$ for more than 1000 frames (due to the presence of shops), *(ii)* some people are visible in $C_2$ for only a few frames (the minimum is 4 frames), and *(iii)* the travelling time of people to go from one camera to the next is highly variable and goes from 7 seconds to 113 seconds (see for example the large difference between person 36 and person 43 in Fig. 3.9). The maximum speed of people ($s_i$) registered in $C_1$ within $T_p = 50$ frames varies from 0.527 units/frame to 1.489 units/frame on $\mathcal{M}$ (with mean 0.867 units/frame and standard deviation 0.192 units/frame). Note that 1.489 units/frame corresponds to a running person. In addition to this, exit regions in $C_1$ present illumination conditions that are more similar to $C_2$ than the entry regions of $C_1$ and people are more likely to be occluded in the exit regions of $C_1$ than in the entry regions of $C_1$ due to the perspective. In order to account for these characteristics of the dataset, we perform re-identification using two different sets. Let us call EN1EN2 the first set where entry regions of $C_1$ are associated with entry regions of $C_2$

[3] Part of the map has been created using information from the London Gatwick airport website http://www.gatwickairport.com/.

[4] For the interested reader, a discussion on CMC curve can be found in Lian *et al.* [64] while alternative measures are introduced in Leung *et al.* [61] and Bäuml and Stiefelhagen [9].



(a) (b)

Figure 3.8: Variations of people walking speeds from Camera 1 ($C_1$) and Camera 3 ($C_2$) at London Gatwick airport [43] calculated on the top-view map. (a) Average speed difference in the two cameras; (b) travelling time to go from $C_1$ to $C_2$.

Table 3.2: Parameters of the proposed Multi-Goal Social Force Model (MG-SFM). $\mathcal{M}$: top-view map; $m_i$: person's mass; $A_B$: weight associated to the barrier force; $B_B$: barrier interaction range; $w$: weight for actual velocity; $|G|$: number of goals, $|\mathcal{K}|$: number of key regions; $T_p$: number of previous frames to calculate actual velocity; $\varepsilon_v$: value for low velocity thresholding; $\varepsilon_g$: number of units for goal-reached thresholding.

| **Parameter** | size($\mathcal{M}$) | $m_i$ | $A_B$ | $B_B$ | $w$ | $\lvert G\rvert$ |
|---|---|---|---|---|---|---|
| **Value** | $577 \times 961$ units | 70 Kg | 60000 N | 1 unit | 0.3 | 8 |

| **Parameter** | $\lvert\mathcal{K}\rvert$ | $\tau_i$ | $T_p$ | $\varepsilon_v$ | $\varepsilon_g$ |
|---|---|---|---|---|---|
| **Value** | 3 | 1 frame | 25 frames | 0.1 | 5 units |

and EX1EN2 the second set where exit regions of $C_1$ are associated with entry regions of $C_2$. A single snapshot for each person is extracted in each region (see also Sec. 3.4).

Table 3.2 summarises the parameters used in the evaluation for the MG-SFM (Sec 3.3.2). $A_B$ is set high and $B_B$ is set to 1 in order to implement barrier avoidance while letting people move in the environment without too much influence. Using Eq. 3.9 it can be seen that the influence of the barriers on a person is negligible at a distance of about 10 units. We consider the mass $m_i$ of each person to have the same value [38] and we set it to 70 Kg [29]. The actual velocity is calculated during the last 1 second of video (25 frames).

For association using appearance (Sec 3.5), we compare the use of the L1-Norm, the Bhattacharyya Distance (BD) and the rankSVM (rankSVM has comparable results to the Ensemble SVM) adopted in [28, 84, 85, 121]. The training for the rankSVM is performed with 60 people's patches from $C_1$ and $C_2$ (this set does not overlap with the testing set).

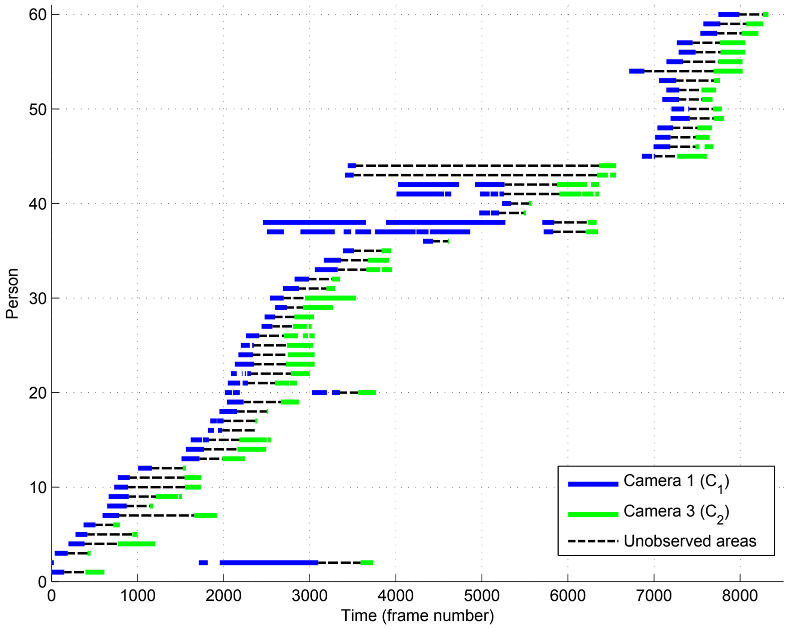
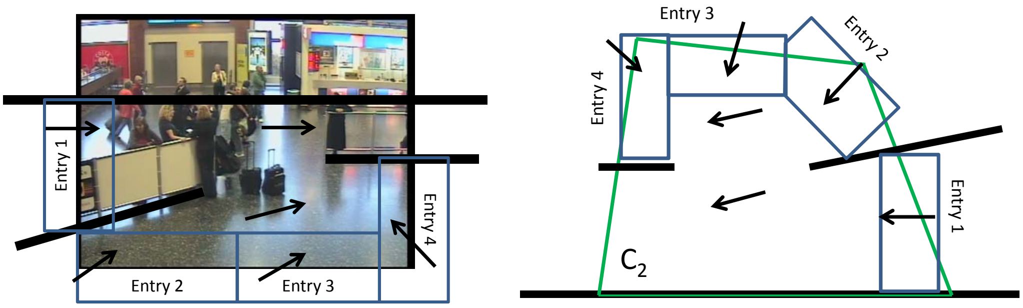


Figure 3.9: Time evolution of people from Camera 1 ($C_1$) to Camera 3 ($C_2$) at London Gatwick airport [43]. Blue line: time elapsed when a person is observed in $C_1$. Dotted line: time elapsed in unobserved regions. Green line: time elapsed when a person is observed in $C_2$.

We compare LBM and MG-SFM as person's motion modelling with two baseline methods for spatio-temporal calibration based on the average travelling time of people to go from $C_1$ to $C_2$. Let TTALL be the first method that calculates the average travelling time of all people that go from $C_1$ to $C_2$, and considers it as the expected travelling time of each person. This method is similar to the one proposed in [16] where people's travelling time is used to make hypotheses for re-identification with the difference that, in our case, training and testing sets are the same (a tough comparison for our method). Let TT4REG be the second method that divides $C_2$ into four entrance regions and calculates the average travelling time of people that only enter the specific region. This creates an expected travelling time for each region. Figure 3.10 shows the four regions, where arrows correspond to possible direction of motion. Note that TT4REG is trained by assuming the region of reappearance of each person as known, thus creating a tough comparison for our method (as for TTALL). For both TTALL and TT4REG, the average travelling time is calculated as mean, but similar results are obtained using the median and a Gaussian distribution around the mean time with standard deviation calculated on data from all people for TTALL and from people in each region for TT4REG. We perform a ranking for person association by calculating the absolute time difference between the time step when a person reappears, and the results of TTALL and TT4REG. Since in LBM and MG-SFM we



(a)

(b)

Figure 3.10: Entry regions of Camera 3 ($C_2$) at London Gatwick airport [43] for the spatio-temporal calibration method TT4REG. TT4REG is employed as a baseline method for motion propagation. In each of the four regions, the expected travelling time is calculated as the mean travelling time of all the people entering in that specific region (see text for details). (a): image plane; (b): top view. Black line: barrier or wall. Green line: FOV of $C_2$. Blue area: possible entry. Black arrow: possible people's movement. Of the 60 people used in Sec. 3.6, 10% enter in Entry 1, 15% in Entry 2, 70% in Entry 3 and 5% in Entry 4.

consider only candidates within a time interval of $\pm\Delta_t$, in order to make a fair comparison we also consider a correct association when a person arrives within a time interval of $\pm\Delta_t$ frames of the expected time. Let us call the corresponding methods as TTALL-50 and TT4REG-50. As already mentioned in Sec. 3.3.2, we set $\Delta_t = 2 \cdot T_p$ for our experiments.

Finally, as we focus on modelling people's movements in unobserved regions and on re-identification, we consider the single-camera detection and tracking task solved by employing annotated heads (see App. A [J1] for a possible solution of this task). Feet locations are estimated starting from the head positions as explained in Sec. 3.4.

We organise the experiments as follows. In Sec. 3.6.2, the MG-SFM is validated for motion prediction, and results using different strategies for calculation of the desired speed are presented and compared to the baseline methods. Section 3.6.3 presents a comparison on motion prediction models (MG-SFM, LBM and baseline methods), and shows that these motion models can create good re-identification candidates and that the best performances are obtained by combining appearance and spatial cues for association. In Sec. 3.6.4, the results obtained using the proposed person's representation (Fig. 3.7) are compared to those obtained using the full body of a person, showing the superior performance of the proposed representation.

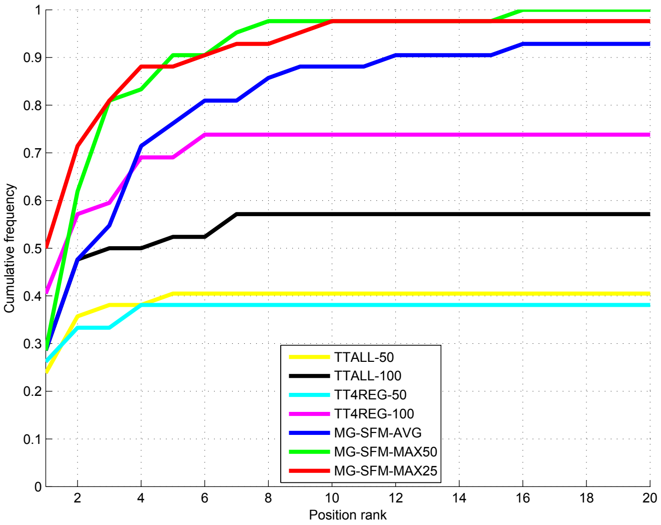


Figure 3.11: Cumulative Matching Characteristic (CMC) curves for person re-identification using different spatio-temporal modellings of people in unobserved regions (no appearance is used). The results are obtained with the MG-SFM where three different strategies are used for desired speed calculation, and with four spatio-temporal calibration methods (see text for details). Dataset: 42 people going from Camera 1 ($C_1$) to Camera 3 ($C_2$) in the i-LIDS dataset from London Gatwick airport [43]. X-axis: person re-identification ranking. Y-axis: frequency accumulation of the correct person re-identification ranking.

### 3.6.2 Validation of the Social Force Model-based motion prediction

This section presents an in-depth analysis of the MG-SFM performance on 42 people[5]. We obtain the re-identification results with the three strategies for desired speed calculation presented in Sec. 3.3.2 (MG-SFM-AVG, MG-SFM-MAX25 and MG-SFM-MAX50) where association is performed by Euclidean distance, and we compare them with TTALL-50 and TT4REG-50. Moreover, we perform a more challenging comparison with TTALL and TT4REG by considering a correct association if a person arrives within a time interval of $2 \cdot \Delta_t$. Let us call the corresponding methods TTALL-100 and TT4REG-100. Note that the ranking of the possible candidates is calculated by only using spatial Euclidean distance and no association method based on appearance of targets has been used, as we are now validating the MG-SFM as a motion model. Figure 3.11 shows the final results by CMC curve.

MG-SFM-MAX25 outperforms the baseline methods based on average travelling time (TTALL and TT4REG), while MG-SFM-MAX50 outperforms them starting at position rank 2. With MG-SFM-MAX25 we obtain 50% of correctly re-identified people, compared to 41% of TT4REG-

[5] We restrict the dataset from 60 to 42 people in order to facilitate the analysis while maintaining its principal characteristics.

100, and 29% of MG-SFM-MAX50 and MG-SFM-AVG. Furthermore, if we consider the first four positions in the ranking we have 88% and 83% of correct re-identifications for MG-SFM-MAX25 and MG-SFM-MAX50, respectively. On the other hand, MG-SFM-MAX25 never reaches 100% in the re-identifications task because the method cannot predict the behaviour of a person who travels at an average speed in $C_1$, and then takes a long time to reappear in $C_2$ (more than 31% of the average travelling time of their reappearance region). In general, MG-SFM-MAX50 and MG-SFM-MAX25 are better at modelling people's desired speed compared to MG-SFM-AVG. In fact, it is likely that the registered highest speed describes well the desired speed that a person would maintain if there would not be any constraints in the environment, and for this reason, in the rest of the thesis, we shall only use MG-SFM-MAX25 and MG-SFM-MAX50.

Finally, Fig. 3.12 shows the confusion matrix obtained with MG-SFM-MAX25, reporting the distances resulting from Eq. 3.12. It is interesting to note that person 12 and person 14 are re-identified with rank 2, and the distance between the best prediction and reappearance position is less than 3 units, hence very close to the correct re-identification (the green circle in Fig. 3.6 is 20 units). Difficult cases for our motion modelling are when people exit $C_1$ at roughly the same position and time step. An example is person 21, 22, and 23. However, since these people exit at different velocities, we can still have rank 2 and 1 for person 21 and 22, respectively, because our model creates different predictions for each of them. A second example is provided by people 37 and 38, who walk and exit together $C_1$ at approximately the same velocity, and reappear in the same region in $C_2$. In this case, the distance between predictions and observed trajectory is less than 7 units between the two and over 81 units from person 42: a wrong hypothesis for the re-identification. Furthermore, only two people (number 18 and 25) have the correct ranking values over 20 units and only one person (number 7) is out of ranking because the predictions are too far away in time. In these latter cases, people substantially vary their speed when unobserved and, even if these variations are not explicitly modelled, our method can cope with them to some extent. These results show how MG-SFM can estimate well people's movement in unobserved regions for the re-identification problem and even in the cases when the method cannot perfectly solve the re-identification problem, it can give reasonable hypotheses on the position and the time of reappearance of a person. It is also important to notice that MG-SFM does not need any training phase for learning common paths that people follow or average travelling times, unlike

TTALL and TT4REG [16].

### 3.6.3 Re-identification results and analysis

In this section, we compare the proposed LBM and MG-SFM using a combination of appearance and spatial association applied to EX1EN2 because the propagation of paths is between exit regions of $C_1$ and entry regions of $C_2$. First, we validate LBM by comparing it with the MG-SFM using only Euclidean distance as an association method, similarly to Sec 3.6.2. Speed in LBM is calculated with $T_p = 25$ frames and $T_p = 50$ frames (namely LBM25 and LBM50, respectively) as for MG-SFM. Figure 3.13(a) shows the CMC curves where the results of the baseline methods TTALL-50 and TT4REG-50 are also reported. As already shown in Sec 3.6.2 for 42 people, poor results are obtained with TTALL-50 and TT4REG-50 due to the high variability of people's travelling times. LBM50 gives better results than LBM25 on average. Moreover, LBM shows results slightly worse or comparable to MG-SFM-25 and MG-SFM-50 for the first three ranking positions, respectively, and better results after ranking three. However, LBM requires less time to be computed and a smaller number of parameters to set, thus resulting in a better applicability of the method.

We now perform re-identification using LBM50 as motion prediction, where for association L1-Norm, BD and rankSVM are used as appearance methods. We compare these results with the results obtained by only using appearance without any spatio-temporal feature to show the improvement that motion prediction can give to re-identification. Figure 3.13(b) shows that the re-identification score is improved by 28% for L1-Norm, 28% for BD and 20% for rankSVM when LBM is employed. These results highlight how LBM creates good candidates for re-identification by restricting the possible candidates to only those close in time and space to the reappearance time and location, respectively. Finally, we apply LBM as motion prediction and we perform association by using a combination of appearance methods (L1-Norm, BD and rankSVM) that evaluate the similarity of people's patches and a spatial method (Euclidean distance) that evaluates the closeness in space of the predictions. In order to select the best combination weights, we test the re-identification performance by applying different weights to BD and Euclidean distance, resulting in the CMC curves in Fig. 3.14(a). Then, we decide to weight 50% the association ranking given by appearance and 50% the one given by Euclidean distance, since it has the highest re-identification score among the possible weights. Figure 3.14(b) shows the CMC curves. The black CMC curve is a baseline result obtained by LBM50 and Euclidean distance where no

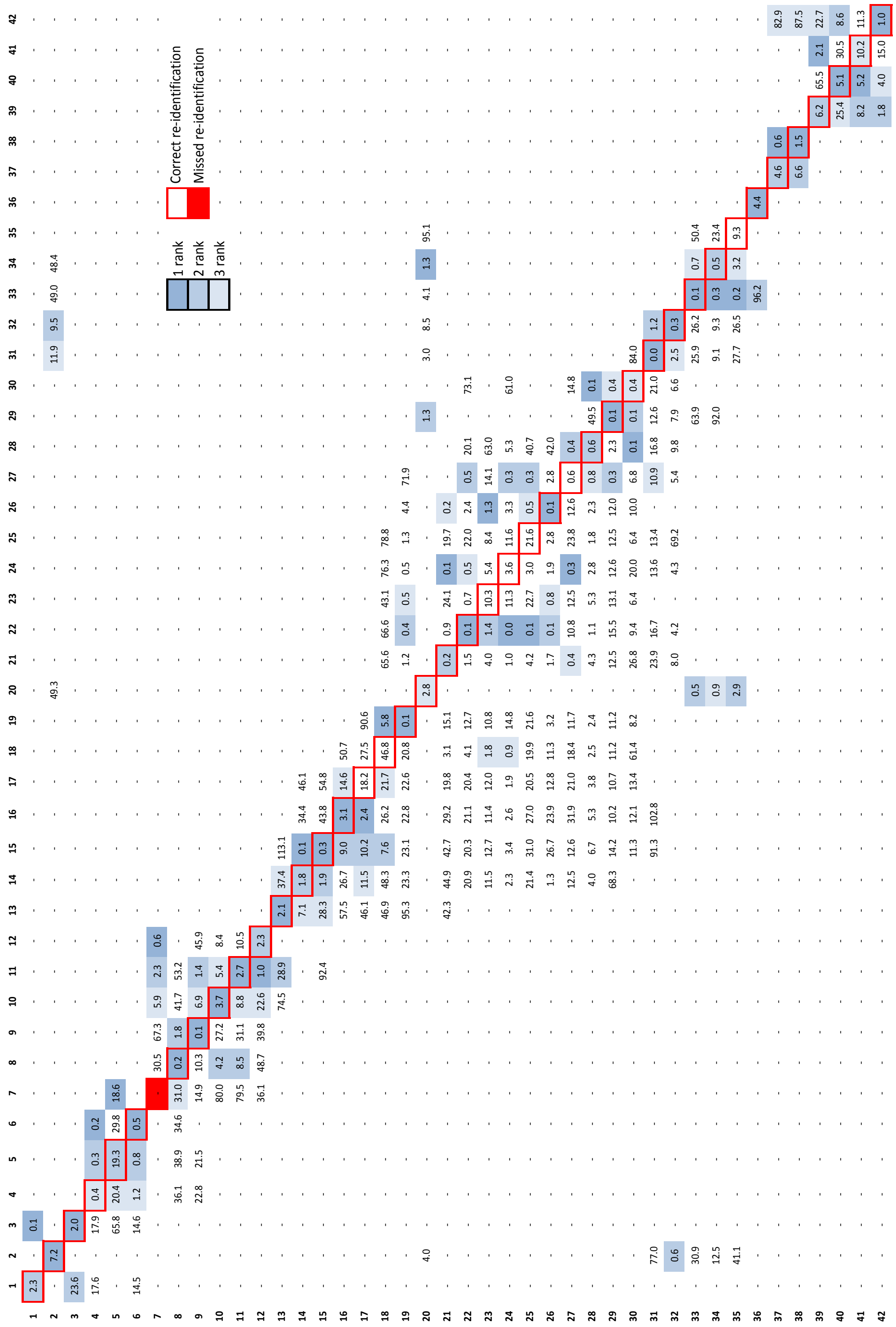
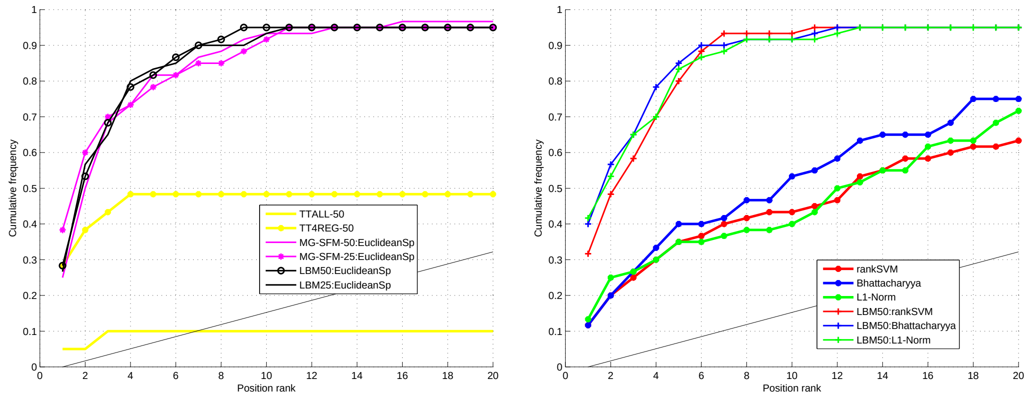


Figure 3.12: Confusion matrix for person re-identification as a result of MG-SFM-MAX25. Each row corresponds to a person $P_i$ to be re-identified. Each column corresponds to possible candidates $P_r$ for re-identification. Each cell contains the minimum distance between the closest predicted trajectory and the trajectory in the observed region (calculated with Eq. 3.12). Red cell: missed re-identification ranking. Coloured cell: different person re-identification ranking. Red-bordered cell: diagonal of the original confusion matrix (in the ideal case it contains the minimum distance). Cell with '-': the predicted trajectory is too far away in time to be considered and therefore removed from the candidate list.



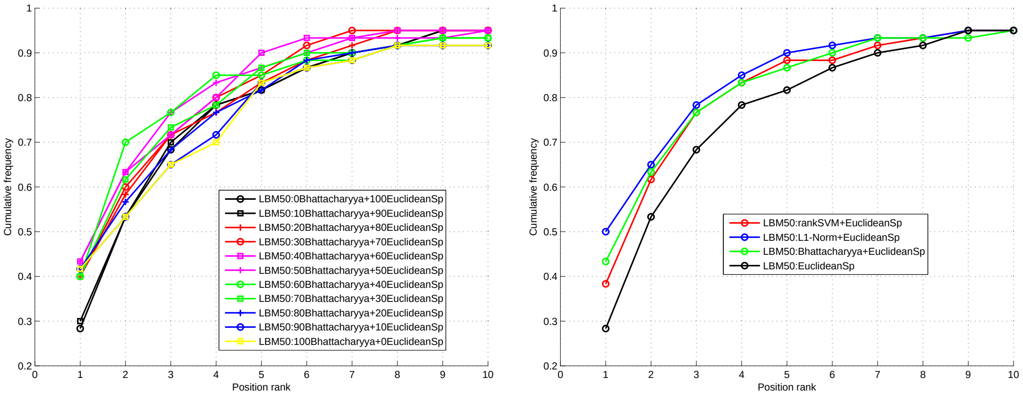
(a)

(b)

Figure 3.13: Cumulative Matching Characteristic (CMC) curves for person re-identification. Dataset: 60 people going from Camera 1 ($C_1$) to Camera 3 ($C_2$) in the i-LIDS dataset from London Gatwick airport [43]. (a) Different spatio-temporal modellings of people in unobserved regions (no appearance features are used): Multi-Goal Social Force Model (MGSFM) and LBM where association is performed using Euclidean distance, and TTALL and TT4REG as baseline methods that use the average travelling time (see text for details). (b) Appearance methods (L1-Norm, Bhattacharyya distance and rankSVM) applied on the full dataset and on the re-identification candidates generated for each person by LBM50.

appearance is used (this curve is the same as in Fig. 3.13(a)). Given the re-identification candidates provided by LBM50, the combination of Euclidean distance with L1-Norm (blue CMC curve), BD (green CMC curve), and rankSVM (red CMC curve) gives 50%, 43%, and 38% for the re-identification score, respectively. These re-identification scores average to 44% that provides an improvement of 6% when compared to the average results obtained by LBM50 as motion prediction with only appearance for association (the three top CMC curves in Fig. 3.13(b)) and an improvement of 16% over those obtained by LBM50 as motion prediction with only Euclidean distance for association (black CMC curve in Fig. 3.14(b)).

Note that the CMC curves reported in the figures are not always monotonically increasing and they remain constant for a few ranks after the first 5-10. This happens because we use a dataset of 60 people and we reach over 80% in the CMC curves within the lower ranks, so the number of people that could be re-identified at higher ranks are only a few. This is desirable for algorithm performance, but constant trends generated at higher ranks highlight the fact that there exist some people that are very difficult to be re-identified by the algorithm. Moreover, in some of the figures the CMC curves never reach 100%, as already mentioned in Sec. 3.6.2 for Fig. 3.11. In fact, the time interval $\Delta_t$ (Sec. 3.3) can sometimes be too restrictive and this results in some correct candidates being left out of the set of possible re-identification candidates. A solution



(a)

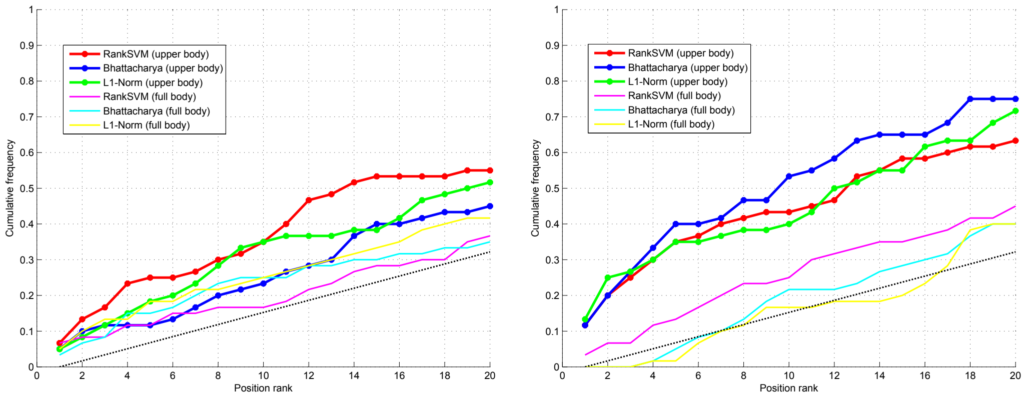
(b)

Figure 3.14: Cumulative Matching Characteristic (CMC) curves for person re-identification using LBM50 to generate re-identification candidates for each person. (a) Association performed by different weighted sums of Bhattacharyya distance (appearance measure) and Euclidean distance (spatial measure). (b) Association performed by only using Euclidean distance (black CMC curve) and by weighted sum (50%-50%) of Euclidean distance and different appearance methods (L1-Norm, Bhattacharyya distance and rankSVM).

for this issue would be to increase the $\Delta_t$ value at the cost of a decrease in the re-identification performance.

### 3.6.4 Validation of the proposed person's representation

In order to evaluate our proposed representation model, we compare the results obtained with a full-body model [28, 84, 85, 121] and those obtained using the shape from Fig. 3.7. In these experiments, only appearance features are used without any motion prediction. Figures 3.15(a) and 3.15(b) show the results by Cumulative Matching Characteristic (CMC) curves. It is possible to notice that the upper-body model is a more suitable shape to use for re-identification than the full body. In particular, since people are more likely to be occluded when they exit $C_1$, Fig. 3.15(b) shows a higher improvement compared to Fig. 3.15(a) that considers the entry of $C_1$. Moreover, the L1-Norm and BD show a considerable improvement from Fig. 3.15(a) (EN1EN2) to Fig. 3.15(b) (EX1EN2) compared to rankSVM that has a more stable behaviour. This is because the rankSVM is a learning-based method with a cross-camera colour calibration implicitly performed in the training phase and hence is more robust to illumination changes.



(a) (b)

Figure 3.15: Cumulative Matching Characteristic (CMC) curves for person re-identification using different spatial supports for feature calculation. Only appearance features are used for association without any motion prediction. Dataset: 60 people going from Camera 1 ($C_1$) to Camera 3 ($C_2$) in the i-LIDS dataset from London Gatwick airport [43]. Full body is the left-most image in Fig. 3.7, upper body is the right-most image in Fig. 3.7. (a) People entering $C_1$ associated with people entering $C_2$. (b) People exiting $C_1$ associated with people entering $C_2$.

## 3.7 Summary

In this chapter, we proposed a person re-identification framework divided in two main phases: *(i)* generation of reappearance candidates of people and *(ii)* association of people across cameras. In order to generate the reappearance candidates, we modelled people's movement in unobserved regions using the top-view map of the environment. We proposed two different models for movement propagation. In the first, crossing regions in the site are marked as landmarks on the top view and people's movement is modelled using a graph-based approach (LBM). In the second, a person's desire to reach specific goals in the site while avoiding obstacles is modelled using the Multi-Goal Social Force Model (MG-SFM), a modification of the Social Force Model commonly used in crowd simulation.

A person that reappears in the next camera is then associated with those candidates within a time window around the person's reappearance time. Association is performed using two measures: *(i)* the spatial Euclidean distance calculated on the top view between the position of the candidates and the position of reappearance of the specific person; and *(ii)* the Bhattacharyya distance between appearance features extracted from the reappeared person and people that have generated the candidates. In order to reduce the presence of background and the probability of occlusion, we extracted appearance features from a patch around the head obtained as reported in Fig. 3.7, thus making the method suitable for crowded scenes. Finally, the association is performed by equally weighting the contribution of spatial Euclidean and Bhattacharyya distances leading to an association ranking.

We used a challenging dataset of 60 people from London Gatwick airport in order to compare the proposed methods with state-of-the-art methods in a realistic crowded scene. Approaches solely based on appearance features extracted from the full body of a person have performances close to random, nevertheless when they are extracted from part of the upper body they can reach 40% in the first 10 ranking positions. LBM and MG-SFM have been compared in the generation of re-identification candidates using spatial association. With the best settings for both LBM and MG-SFM, MG-SFM performs better for the first three positions of the re-identification ranking while LBM presents better results after the third position. Overall, by employing LBM for the generation of re-identification candidates and different strategies for association across cameras, the re-identification score (rank 1) can reach 41.67% when only appearance is used, 28.33% when only motion is used, and 50% when their rankings are summed with equal weights. In general, similarly to what has been reported in the state of the art [47, 65], it is important to highlight the fact that spatio-temporal cues, when available, should be used in combination with appearance methods for re-identification, as this combination normally outperforms methods solely based on spatio-temporal features or on appearance features.

# Chapter 4

# Human interaction analysis

## 4.1 Introduction

About 50-70% of human walking activity takes place in groups [105]: video monitoring of spatially interacting humans is therefore very important for analysing people's behaviours [95, 122]. In this chapter, our aim is to detect those people that know each other and form a group, and those that interact with each other for short periods of time. We concentrate on scenarios of low- and mid-density crowds, and we find interacting people by analysing people's trajectories and by calculating the expected people's movements. The task becomes very challenging when ambiguous situations are created by people that stand or pass very close to each other without interacting. In order to address these situations, we extend the Social Force Model with latency method for group detection presented in Šochman and Hogg [105] by defining plausible human behaviours for the localisation of group formations. With the aim of enabling group detection in situations that were not previously possible, we improve the model by incorporating relationship constraints such as walking in the same direction and decelerating when approaching individuals who are standing still (Sec. 4.2). Moreover, we track each centre of interaction over time with a graph-based tracker to enforce the spatio-temporal consistency of the detections (Sec. 4.3). Figure 4.1 shows the block diagram of the overall solution. We validate the proposed framework on three different datasets and show that it outperforms existing methods using the one minus False Positive rate (1-FP) and the Group Detection Success Rate (GDSR) [10] (Sec. 4.4).

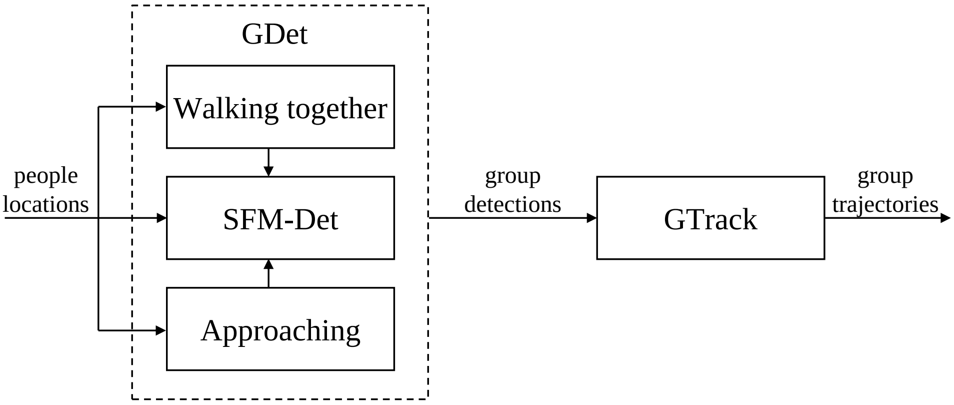


Figure 4.1: Block diagram of the proposed approach (GDet) for group detection and tracking, showing the embedding of walking together, approaching and SFM-Det (Social Force Model group Detection [105]). GTrack corresponds to Group Tracking.

## 4.2 Interaction detection

Let $\mathcal{P} = \{P_1, P_2, \ldots, P_N\}$ be the set of $N$ people walking or standing still in the monitored scene, and $\mathbf{p}_i(t) = (x_i(t), y_i(t))$ be the feet position of person $P_i$ at time $t$ on the rectified image[1]. Šochman and Hogg [105] proposed to detect interactions among people by analysing instantaneous forces with a Social Force Model (SFM-Det); however, this method fails to detect groups in situations when the paths of walking people cross over each other or when people pass near a stationary group (Fig. 4.2). In our approach, we specifically model these situations. Note that, compared to Sec 3.3.2 where SFM is used as a motion model to propagate people's movement in unobserved regions, SFM-Det analyses the observed trajectories in order to describe the expected movement of each person and to find which people are in group.

In the SFM-Det, the desired velocity of a person is defined as the average velocity observed in the time interval $[t, t+T_p]$, and the force $\mathbf{f}_{D_i}(t)$ as the displacement between actual velocity and desired velocity. Let us define the set of people $\mathcal{H}_i = \mathcal{P} \setminus P_i$ and the contour of non-walkable areas $B$ (walls and barriers), where $B$ are the spatial locations defined by a vector $(x_B, y_B)$ for the horizontal and vertical coordinates, respectively. The sum of repulsive forces, $\mathbf{f}_{\mathcal{H}_i}(t)$ and $\mathbf{f}_{B_i}(t)$, is inversely proportional to the distance between $\mathbf{p}_i(t)$ and $\mathcal{H}_i$, and $\mathbf{p}_i(t)$ and $B$ in order for $P_i$ to be at a comfortable distance from other people $\mathcal{H}_i$ and from walls and barriers $B$, respectively. In particular,

$$\mathbf{f}_{\mathcal{H}_i}(t) = \sum_{j|P_j \in \mathcal{H}_i} \mathbf{f}_{ij}(t), \tag{4.1}$$

[1] Note that, compared to Ch. 3, we do not need a top view because we now deal with single cameras.

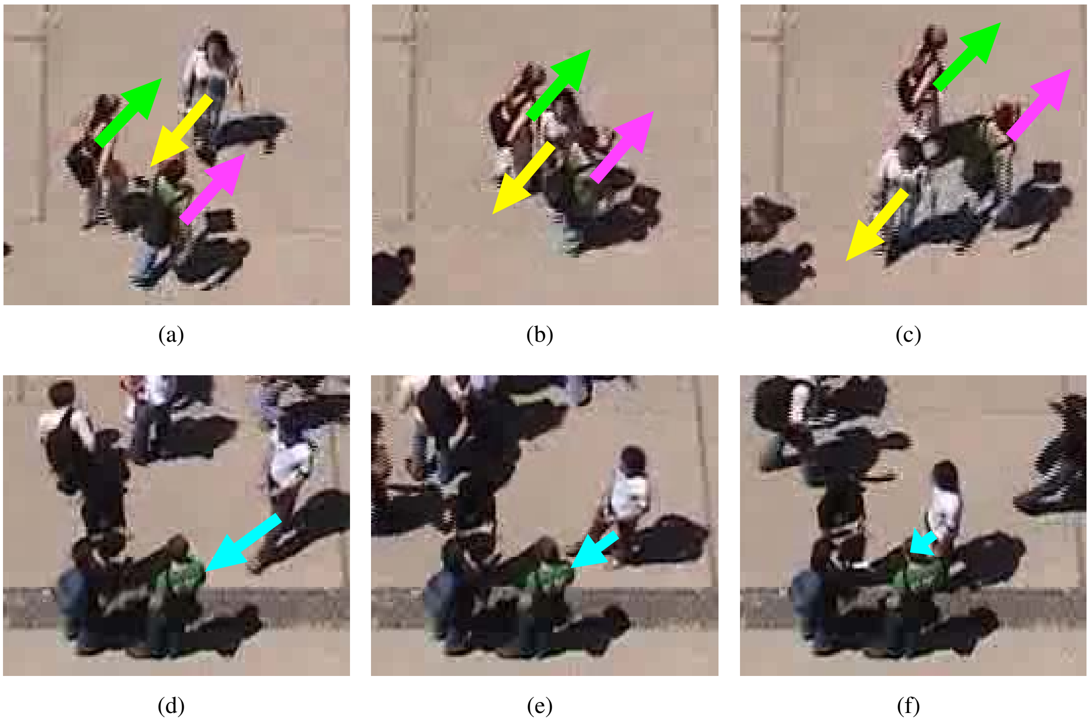

(a) (b) (c) (d) (e) (f)

Figure 4.2: Visualisation of plausible human behaviour constraints. (a)-(c) A person crossing a group of two people walking together is not detected as part of the group; (d)-(f) A person approaching a stationary group is detected as a member only if he/she decelerates and stops in proximity of the group.

where $\mathbf{f}_{ij}(t)$ indicates the interaction force generated by $P_j$ on $P_i$ and

$$\mathbf{f}_{B_i}(t) = \sum_{b \in B} \mathbf{f}_{ib}(t), \tag{4.2}$$

where $\mathbf{f}_{ib}(t)$ indicates the barrier force generated between a barrier $b$ and person $P_i$. In our implementation, we model the module of $\mathbf{f}_{ij}(t)$ and $\mathbf{f}_{ib}(t)$ as a decreasing exponential in the direction of the interaction, similarly to previous works [38, 39, 59, 66]. Alternative models include a circular exponential where person $P_i$ is influenced by all the surroundings [50, 93] and an elliptical exponential that reshapes based on the pedestrian speed [50]; the exponential function has also been demonstrated to be the best to model interaction forces [74]. Moreover, in the SFM-Det, the attractive force, $\mathbf{f}_{\Gamma_i^\gamma}(t)$, is the force that keeps people within the same group $\Gamma_i^\gamma(t) \subseteq \mathcal{P}$ [105], where $P_i \in \Gamma_i^\gamma(t)$ as in Sec. 1.3[2]. The sum of forces describes the variation of people's movement

[2] In the rest of the chapter, we drop the index $\gamma$ for better readability.

that, at instant $t$ and for a person $P_i$, results in

$$m_i \frac{d\mathbf{v}_i(t)}{dt} = \mathbf{f}_{D_i}(t) + \mathbf{f}_{\mathcal{H}_i}(t) + \mathbf{f}_{B_i}(t) + \mathbf{f}_{\Gamma_i}(t), \tag{4.3}$$

where $\frac{d\mathbf{v}_i(t)}{dt}$ is the actual variation of the velocity over time and, unlike Eq. 3.6, the mass of people, $m_i$, does not need to be set to a specific value because it is implicitly determined by people's trajectories from the right-hand side of Eq. 4.3. The left-hand side of Eq. 4.3 models the variation in space of a person's movement at each time instant and the position of each person is then predicted using

$$\mathbf{p}_i^*(t+T_p) = \mathbf{p}_i(t) + T_p\Big(\frac{d\mathbf{v}_i(t)}{dt}\tau + \overline{\mathbf{v}}_i(t)\Big), \tag{4.4}$$

where $\mathbf{p}_i^*$ indicates the expected position of $P_i$ determined using the SFM, $\tau$ is the time interval during which the Eq. 4.3 is calculated (we set $\tau = 1$ similarly to Eq. 3.7), and $\overline{\mathbf{v}}_i(t)$ is the actual velocity obtained as the average velocity within the time window $[t - T_p, t]$. Lower values of $T_p$ would result in a model that is sensitive to noise, while higher values of $T_p$ could not reliably describe abrupt velocity variations. Note that, unlike in the Eq. 3.7 where the temporal smoothing is performed with a weighted sum, in the Eq. 4.4 the temporal smoothing is implicit in the calculation of $\overline{\mathbf{v}}_i(t)$ and $\mathbf{p}_i^*(t+T_p)$. In summary, the SFM-Det describes people's movements using Eq. 4.3 that models people's desire to reach and maintain a specific speed and direction, and the fact that people want to move or stand at a comfortable distance from other people and barriers; Eq. 4.4 is then used by the SFM-Det to smooth speed and direction of motion over time by assuming that abrupt changes are unlikely to happen.

Compared to the original SFM [38], the term $\mathbf{f}_{\Gamma_i}(t)$ in Eq. 4.3 is introduced for group behaviour modelling [74, 75] that is used for interaction detection in an iterative algorithm [105]. This algorithm evaluates whether the prediction using Eq. 4.4 commits more error, $\delta$, when there are no group forces involved ($\mathbf{f}_{\Gamma_i}(t) = [0\ 0]$), or when there is an active group force keeping people together and inhibiting them from being repulsed by each other. In Šochman and Hogg [105], all people can interact with each other (within a certain gating radius) in order to form a group, thus resulting in a reliable group detection only in specific scenarios where people do not walk or stand next to each other. However, in crowded cases like the one presented in Fig. 4.3 where people can cross existing groups and in Fig. 4.4 where people walk very close to other people but without stopping, this method may fail. In order to specifically address these challenging sit-

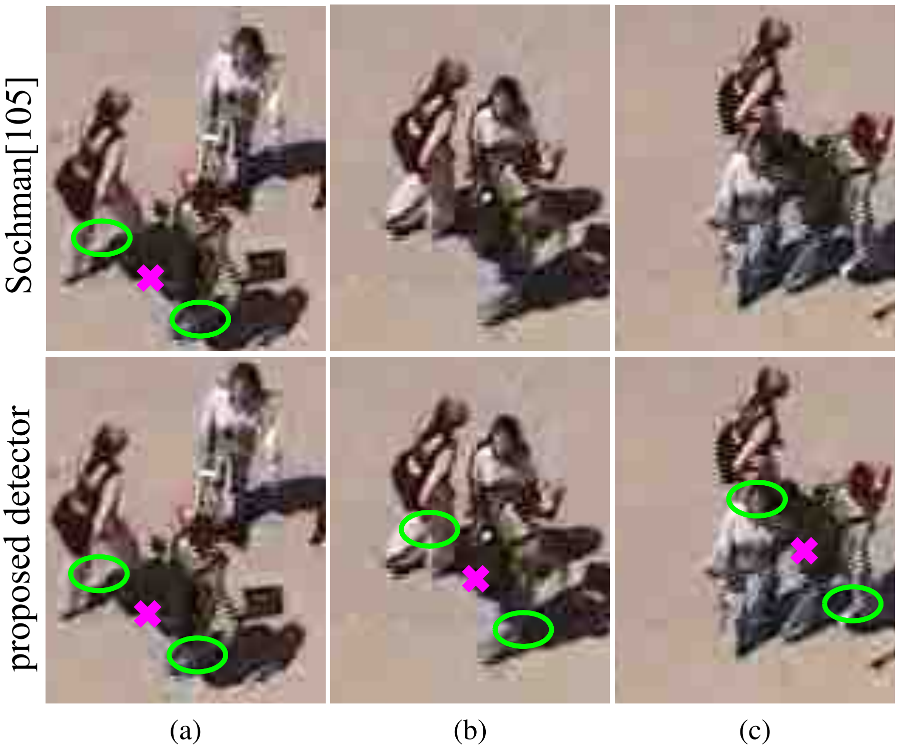


Figure 4.3: Sample of group detection results in the case of a person crossing a group of people walking together. Green ellipse: person (feet location) belonging to the group; Magenta cross: group centroid. From (a) to (c): temporal evolution.

uations, we restrict the set of potentially interacting people to those walking in similar directions (*walking together*) and to those decelerating when approaching a stationary group (*approaching*). The former model defines that only people walking in the same direction can interact (Sec 4.2.1), while the latter model defines that people approaching a stationary group shall decelerate and almost stop in order to be considered as interacting with that group (Sec 4.2.2). We define our model as GDet. Let us call $\overline{\mathcal{H}}_i(t) \subseteq \mathcal{H}_i$ and $\widehat{\mathcal{H}}_i(t) \subseteq \mathcal{H}_i$ the sets of people potentially interacting with person $P_i$ at time $t$, and selected by walking together and approaching, respectively. We then restrict the interactions of $P_i$ to the set $\mathcal{H}_i^*(t) = \overline{\mathcal{H}}_i(t) \cup \widehat{\mathcal{H}}_i(t)$, where $\mathcal{H}_i^*(t) \subseteq \mathcal{H}_i$. In order to consider $\mathcal{H}_i^*(t)$ instead of $\mathcal{H}_i(t)$, Eq. 4.3 is modified by changing the actual variation of the velocity from $\frac{d\mathbf{v}_i(t)}{dt}$ to $\frac{d\mathbf{v}_i^*(t)}{dt}$ and the interaction force from $\mathbf{f}_{\mathcal{H}_i}(t)$ to $\mathbf{f}_{\mathcal{H}_i^*}(t)$ that results in

$$m_i \frac{d\mathbf{v}_i^*(t)}{dt} = \mathbf{f}_{D_i}(t) + \mathbf{f}_{\mathcal{H}_i^*}(t) + \mathbf{f}_{B_i}(t) + \mathbf{f}_{\Gamma_i}(t), \tag{4.5}$$

where the group force $\mathbf{f}_{\Gamma_i}(t)$ is only applied between $P_i$ and $\mathcal{H}_i^*(t)$ at each time $t$. Similarly, in Eq. 4.4 we change the prediction of person $P_i$ position from $\mathbf{p}_i^*(t+T_p)$ to $\overline{\mathbf{p}}_i^*(t+T_p)$ that results

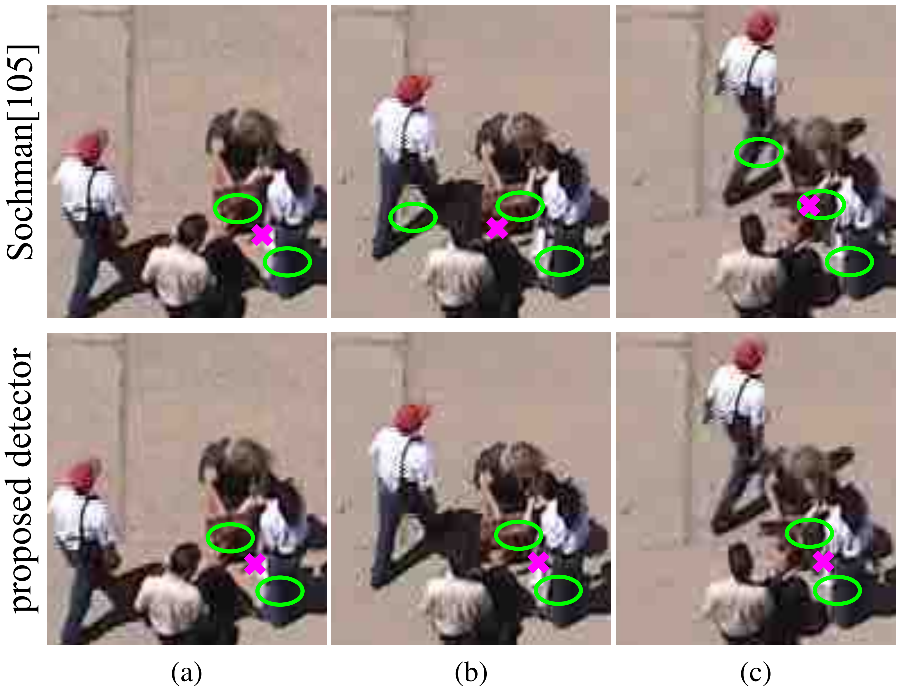


Figure 4.4: Sample of group detection results in the case of a person passing nearby a group. Green ellipse: person (feet location) belonging to the group; Magenta cross: group centroid. From (a) to (c): temporal evolution.

in

$$\overline{\mathbf{p}}_i^*(t+T_p) = \mathbf{p}_i(t) + T_p\left(\frac{d\mathbf{v}_i^*(t)}{dt}\tau + \overline{\mathbf{v}}_i(t)\right). \tag{4.6}$$

Walking together and approaching models are detailed in Sec. 4.2.1 and Sec. 4.2.2, respectively.

### 4.2.1 Walking together model

Let us consider two people, $P_1$ and $P_2$, interacting with each other or with other people, and walking in the same direction within a range of 180° . If a third person $P_3$ walks close to $P_1$ and $P_2$ in the opposite direction, the interaction model SFM-Det erroneously classifies $P_1$ and $P_2$ as non-interacting people (Fig. 4.3) because the movement of $P_3$ interferes in the group detection phase. This problem can be addressed by modelling the walking together that allows people interacting with other subjects only if their direction of motion is coherent within a range of 180°.

Let us consider two groups of people ($P_1$ with $P_2$, and $P_4$ with $P_5$) that walk in the same direction and are far apart enough to not be detected as a single group, while a fifth person ($P_3$) walks in the opposite direction. With the modelling of [105], when $P_3$ is nearby $P_2$ and $P_4$, the repulsive forces $\mathbf{f}_{23}(t)$ and $\mathbf{f}_{43}(t)$ act on their masses (see Eq. 4.1), and when people are almost

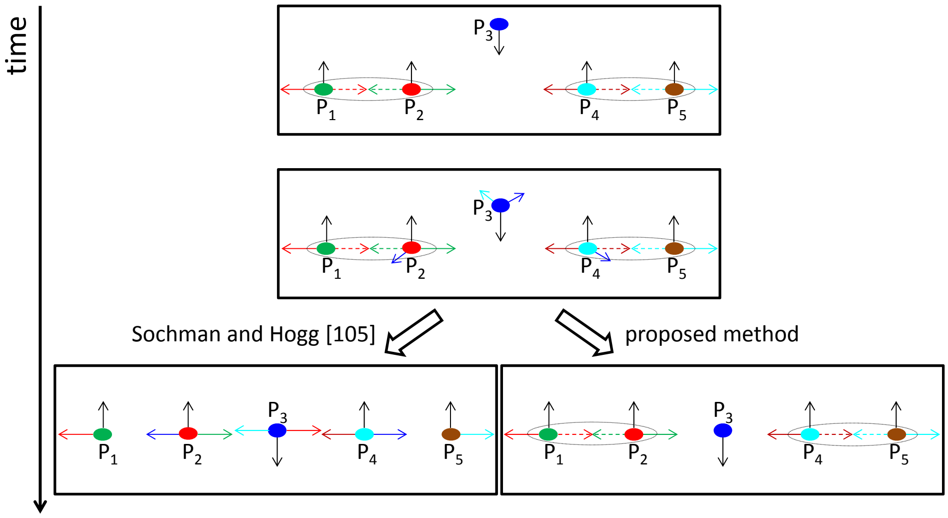


Figure 4.5: Difference between the proposed approach and Šochman and Hogg [105] in how the forces act on people while crossing other groups of people walking together. Coloured ellipses represent people ($\mathcal{P} = \{P_i\}_{i=1}^{5}$); the black arrow is the vector of movement ($\mathbf{f}_{D_i}(t)$); solid and coloured arrows are the repulsive forces ($\mathbf{f}_{\mathcal{H}_i}(t)$); dotted arrows are the attractive forces that form the group ($\mathbf{f}_{\Gamma_i}(t)$); and the dotted ellipses are the detected groups.

aligned, they cancel out with $\mathbf{f}_{21}(t)$ and $\mathbf{f}_{45}(t)$, respectively, thus obtaining

$$\begin{aligned}\mathbf{f}_{21}(t) &\cong -\mathbf{f}_{23}(t)\\ \mathbf{f}_{45}(t) &\cong -\mathbf{f}_{43}(t),\end{aligned} \tag{4.7}$$

a common situation in crowded scenarios (Fig. 4.2). Accepted predictions for the movements of $P_2$ and $P_4$ (Eq. 4.4) are obtained without any group forces, thus leading to a missed detection of the two groups (groups are detected only in the presence of a group force). With the inclusion of the walking together constraint, $P_3$ does not influence $P_2$ and $P_4$ movements with a repulsive force since their motion direction is opposite, and accepted predictions are generated by including the group forces between $P_1$ and $P_2$, and $P_4$ and $P_5$ in the Eq. 4.6, thus leading to correct group detections. Figure 4.5 reports a schematic representation of the forces and shows how group detection is improved by using walking together compared to the original formulation of Šochman and Hogg [105].

### 4.2.2 Approaching model

When dealing with crowded scenes where frequent meetings occur, a single person $P_i$ may interact with a stationary group $\mathcal{S} \subseteq \mathcal{H}_i$, or may pass very close to it in order to shorten the path to

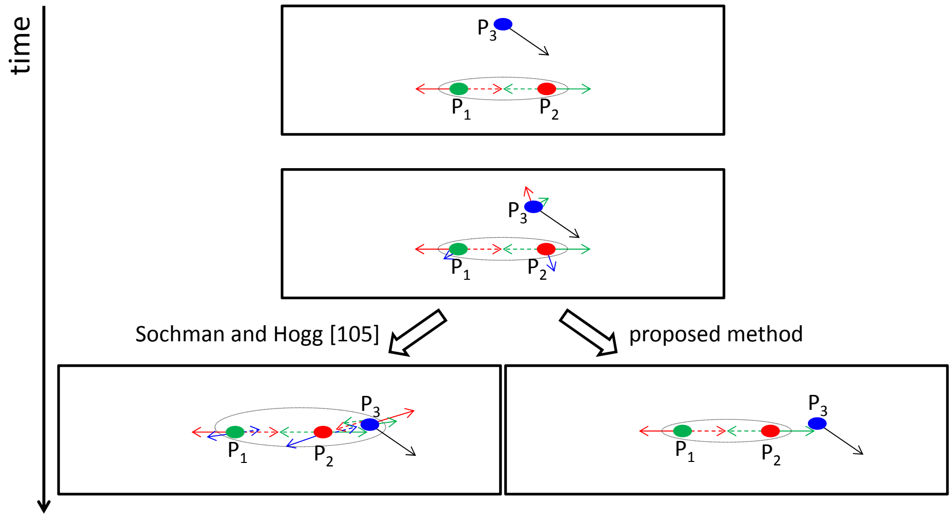


Figure 4.6: Difference between the proposed approach and Šochman and Hogg [105] in how the forces act on people while approaching. Coloured circles represent people ($\mathcal{P} = \{P_i\}_{i=1}^{5}$); the black arrow is the vector of movement ($\mathbf{f}_{D_i}(t)$); solid and coloured arrows are the repulsive forces ($\mathbf{f}_{\mathcal{H}_i}(t)$); dotted arrows are the attractive forces that form the group ($\mathbf{f}_{\Gamma_i}(t)$); and the dotted ellipses are the detected groups.

reach his/her goal. The vicinity of $P_i$ to $\mathcal{S}$ generates a set of high repulsive forces, $\mathbf{f}_{\mathcal{S}_i}$, that in the SFM-Det [105] results in spurious group detections. In order to address this problem, we restrict the possible people interacting with $\mathcal{S}$ to only those decelerating in proximity of $\mathcal{S}$ and stopping within $\widehat{t}$ frames. Figure 4.6 shows the SFM forces with and without considering the approaching model. Initially, when $P_3$ starts to get close to $P_1$ and $P_2$ the repulsive forces start acting on the masses, but until a certain distance is maintained these forces are negligible and do not affect the group force. When $P_3$ is close to $P_1$ and $P_2$, SFM-Det allows the repulsive forces generated by $P_3$ on $P_1$ ($\mathbf{f}_{13}(t)$) and on $P_2$ ($\mathbf{f}_{23}(t)$) to influence their movement predictions and, in order to have accepted predictions, $\mathbf{f}_{13}(t)$ and $\mathbf{f}_{23}(t)$ have to be balanced by group forces that make $P_1$, $P_2$ and $P_3$ being detected as in the same group, thus resulting in a false positive group detection for $P_3$. However, this problem is solved by including the approaching constraint where the interaction of $P_3$ with $P_1$ and $P_2$ is not allowed unless $P_3$ decelerates in proximity of $P_1$ and $P_2$, as if a meeting was about to happen. Figure 4.4 shows a real example of the benefit of using the approaching model for group detection.

## 4.3 Group tracking

After frame-by-frame interaction localisation is performed (GDet), the centres of interaction are tracked in order to enforce their temporal consistency. We define the group locations (or centres of interaction) as the centroid of the positions of the people that form each group. The centres of interaction are associated over time by using a buffered greedy graph-based multi-target tracker [82] (GTrack). In GTrack, the work Poiesi and Cavallaro [82] is adapted to group tracking where the velocity is also included in the algorithm, as mentioned below. At first, short tracks are generated by associating consecutive centroids with Hungarian algorithm (HA)[3]. The association cost used by HA is calculated with the $\ell$-2 norm on the 2D positions of the centroids. Longer tracks are subsequently extracted using GTrack that pair-wise matches short tracks until no alternative better pairings are found. GTrack determines the affinities among the short tracks using position and velocity information, and the association process is performed within a short temporal buffer that involves a sliding window of $\Theta$ frames overlapping for $\theta$ frames. When short tracks are associated, the missing centroids within the temporal gap between the last location of an earlier short track and the initial location of a later one are generated by 2D interpolation. The people forming the group of the earlier short track are propagated up to the later short track.

In Fig. 4.7, we can see the effectiveness of the association of the centres of interaction. A group of two people (light-blue track under white arrow) is passing nearby another group (brown) and, initially, the detector correctly localises the centres of interactions (groups) (Fig. 4.7(a)). When the light-blue group is closer to the brown group (Fig. 4.7(b,c)) the detector fails and assigns the people of the light-blue group to the brown group. However, the tracker manages to recover this erroneous assignment (Fig. 4.7(d)) and returns to tracking the two people belonging to the light-blue group.

## 4.4 Results and analysis

### 4.4.1 Experimental setup

In order to validate the method for localising groups, we compare it with the methods of Šochman and Hogg [105], Bazzani *et al.* [10] and Zanotto *et al.* [118] using BIWI-ETH [81], BIWI-HOTEL [81], and Student003 [105] datasets. BIWI-ETH contains people mainly walking in and

[3] `http://csclab.murraystate.edu/bob.pilgrim/445/munkres.html`, last accessed: March 2013.

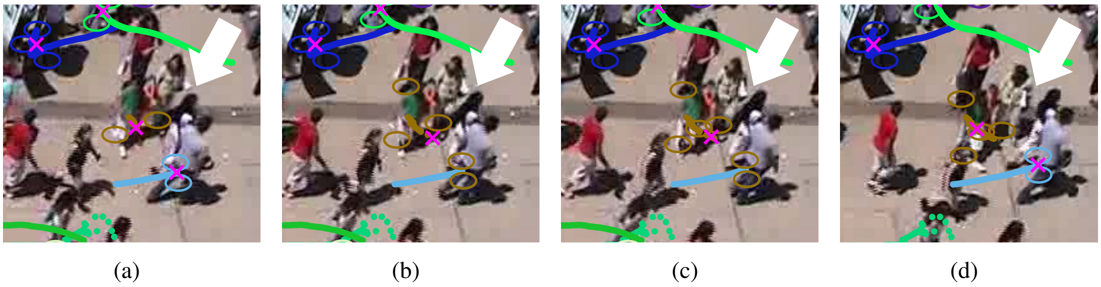


Figure 4.7: Example of track recovery of a two-person group (light-blue track). Coloured circle: affiliation of the people to a group; Coloured line: trajectory; Magenta cross: group detection. (a) Light-blue group is correctly detected and tracked; (b)-(c) for some subsequent frames the light-blue group is erroneously detected as part of a neighbouring group; (d) the light-blue group is correctly recovered by the tracker.

Table 4.1: Details of the datasets used in the experiments. Key - ppg: people per group.

| | **BIWI-ETH** | **BIWI-HOTEL** | **Student003** |
|---|---|---|---|
| **Total number of people** | 360 | 390 | 434 |
| **Number of groups** | 74 | 59 | 109 |
| **Min number of ppg** | 2 | 2 | 2 |
| **Max number of ppg** | 6 | 4 | 6 |
| **Mean number of ppg** | 2.6 | 2.1 | 2.3 |
| **Median number of ppg** | 2 | 2 | 2 |
| **Frame size (pixels)** | 640×480 | 720×576 | 720×576 |
| **Frames per second** | 25 | 25 | 25 |

out of a building; BIWI-HOTEL has more complex movement of people because of the presence of a tram stop and of various barriers; Student003 presents a challenging scenario where people walk in unpredictable directions and get very close to each other. In these datasets, different types of groups are formed, ranging from those in motion to those standing still. Table 4.1 provides additional information about each dataset. The method of Bazzani *et al.* [10] is based on an online DPF for Joint Individual-Group Tracking (JIGT), which is characterised by two conditionally dependent subspaces used to model people's motion and group formations, respectively. Instead, in Zanotto *et al.* [118], observations of people's locations and velocities are generated using a tracker, and group detection is performed online by modelling groups as infinite mixtures solved using DPMM.

### 4.4.2 Validation of group detection and tracking

For group detection (GDet), we use the same parameter setting of Šochman and Hogg [105], except for $\delta$ that we set to a minimum value of 3 (instead of 0) in order to remove noisy group detections. In particular, we set $T_p = \frac{10}{\text{fps}}$ seconds where fps indicates the frames-per-second of the specific dataset used. A person is considered to be standing still when his/her speed is on average less than their shoulder radius - 0.35 meters [105] - within a one-second time window and $\widehat{t} = T_p$ for the approaching model. Group tracking (GTrack) is performed on temporal windows of $\Theta = 25$ frames with a 20% overlap. Like [10, 105, 118], we consider the single-camera person tracking task solved. We compare the results of GDet and the algorithm proposed in [105] without the offline decision (SFM-Det), and those of GTrack applied to the output of GDet and of SFM-Det (let us call them GTrack and SFM-TR, respectively). For JIGT [10] and DPMM [118], we provide the results from the related papers. The evaluation is performed with the mean of all frames of one minus the False Positive rate (1-FP) that indicates the percentage of correct detections of people not belonging to any group and the Group Detection Success Rate (GDSR) metric which calculates the rate of correctly detected groups [10]. A correct detection for GDSR is a group that contains at least 60% of the members annotated in the Ground Truth [10]. Table 4.2 reports the quantitative evaluation, where the results for JIGT and DPMM are only available for the BIWI-ETH dataset. The performance of GDet and GTrack are superior to that of other methods, a part from a small decrement of 1-FP in the Student003 dataset, thus proving the effectiveness of the proposed model for group localisation (Sec. 4.2). Compared to SFM-Det [105], 1-FP of GDet is about the same in all datasets (improvement by 1% in BIWI-HOTEL, decrease by 2% in Student003 and the same value in BIWI-ETH), while GDSR of GDet in BIWI-HOTEL and Student003 is dramatically improved (by 11% and 13%, respectively) since the scene contains people standing still and groups are formed next to each other. In BIWI-ETH, the GDSR improvement of GDet is minor (1%) because the scene is less crowded and groups are located relatively far from each other. Moreover, we believe that the difference between our results (GDet) and those obtained with DPMM for both 1-FP and GDSR (26% and 15%, respectively), are due to the fact that this work is designed for and is therefore more suitable for detecting people switching groups, and because DPMM is an online method while GDet has a small latency.

The group tracker (GTrack) improves the GDSR of GDet by 2% and 1% in BIWI-ETH and

Table 4.2: Result comparison on the BIWI-ETH, BIWI-HOTEL and Student003 datasets using [10] (a) 1-FP and (b) GDSR. GDet: proposed group detection; GTrack: proposed group tracking on the output of GDet; SFM-Det: group detection of Šochman and Hogg [105]; SFM-TR: proposed group tracking on the output of SFM-Det; JIGT: group detection and tracking of Bazzani *et al.* [10]; DPMM: group detection of Zanotto *et al.* [118].

(a) 1-FP

| **Dataset** | **GDet** | **GTrack** | **SFM-Det** | **SFM-TR** | **JIGT** | **DPMM** |
|---|---|---|---|---|---|---|
| BIWI-ETH | 98% | 98% | 98% | 98% | 54% | 72% |
| BIWI-HOTEL | 91% | 91% | 90% | 89% | - | - |
| Student003 | 80% | 80% | 82% | 81% | - | - |

(b) GDSR

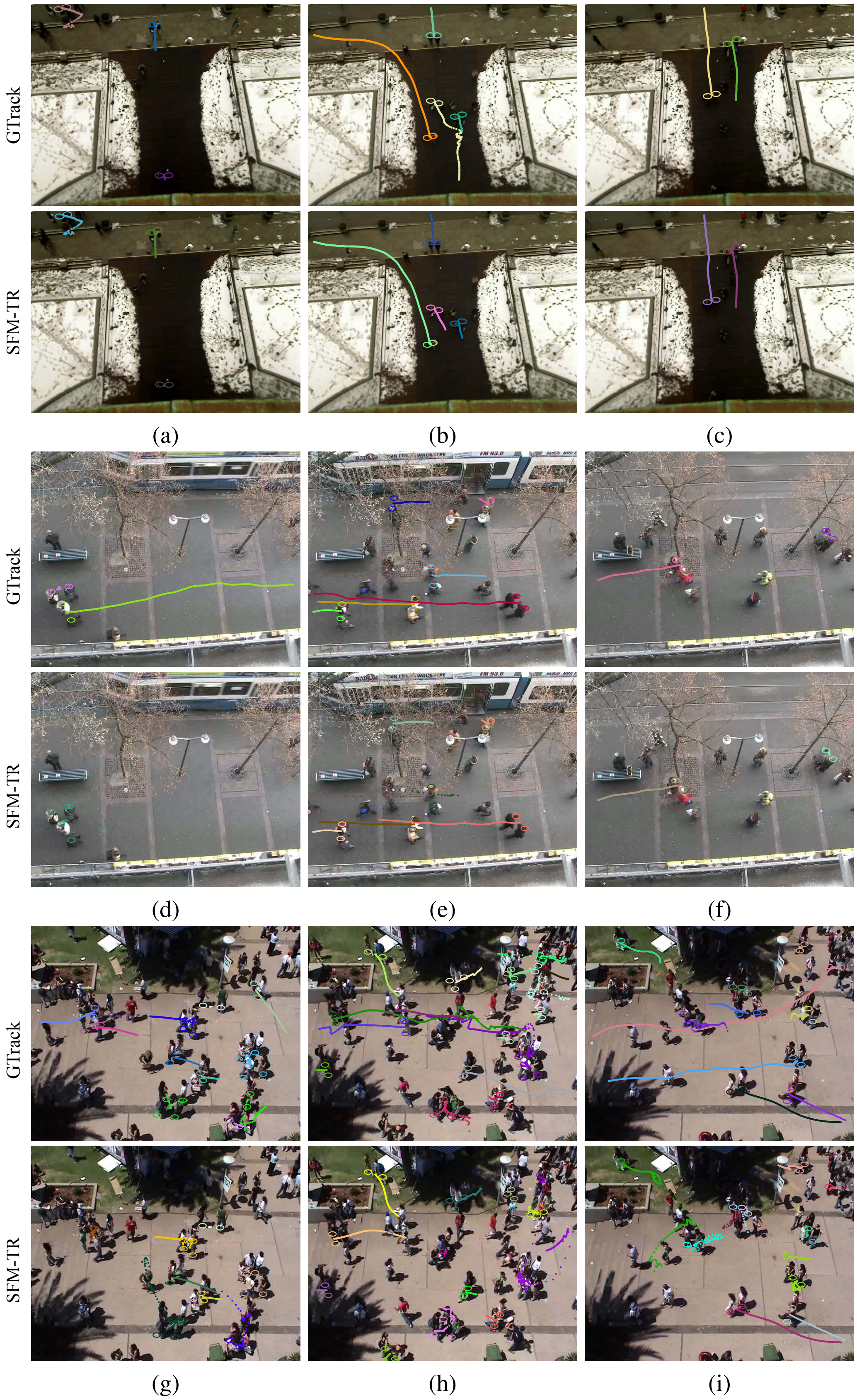
| **Dataset** | **GDet** | **GTrack** | **SFM-Det** | **SFM-TR** | **JIGT** | **DPMM** |
|---|---|---|---|---|---|---|
| BIWI-ETH | 78% | 80% | 77% | 78% | 54% | 63% |
| BIWI-HOTEL | 89% | 89% | 78% | 81% | - | - |
| Student003 | 71% | 72% | 58% | 60% | - | - |

Student003, respectively, whereas GDSR in BIWI-HOTEL remains the same. This is due to the high performance of GDet in BIWI-HOTEL where groups are constantly detected over time, unlike in BIWI-ETH and Student003 where GDet provides less consistent input to the tracker that can then link the centres of interaction over time. 1-FP for GTrack and GDet remains the same for all datasets. GTrack also improves the GDSR of SFM-Det (SFM-TR) by 1%, 3% and 2% in the three datasets, respectively, while maintaining the same 1-FP in BIWI-ETH and decreasing it by only 1% in the other two datasets. Compared to JIGT, GTrack performs 44% and 26% better for 1-FP and GDSR in BIWI-ETH, respectively, because JIGT is more suitable to detect switchings of groups and is online, like DPMM. Figure 4.8 shows the qualitative results as comparison between GTrack and SFM-TR, and how the group tracker links and correctly tracks most of the groups in the scene. From the results shown in Fig. 4.8, we can see some of the challenging situations where the proposed modelling is effective. GTrack has better performance compared to SFM-TR, for example in Fig. 4.8(d) where a stationary group is correctly detected and tracked, even when another one passes close to it. Likewise, in Fig. 4.8(g) on the left, two groups (light blue and purple) that cross each other are consistently localised. In some situations, the group localisation may fail. For example in Fig. 4.8(h), the three people in the middle of the frame are not localised as part of the same group because they just joined together and their group formation is highly unstable, that is they keep moving apart and joining back together while walking, as well as passing through other groups of people. On the other hand, SFM-TR in this case can correctly localise part of the group (two out of three people), even for only few frames until when the group crosses another one coming from the opposite direction.

The overall computational complexity for SFM-Det [105] has an upper limit of $O(N^3)$ where N is the number of people. This is obtained as a multiplication between $O(N)$ operations, that check whether or not people belong to a group, and $O(N^2)$ operations that calculate the set of interaction forces between each pair of people. The GDet method does not increase the computational complexity of SFM-Det ($O(N^3)$) because walking together and approaching models in GDet need $O(1)$ to create the set $\mathcal{H}_i^*(t)$ (Eq. 4.5). Moreover, the implementation of GDet is usually faster than the one of SFM-Det since interaction forces are normally calculated on a subset of people (in most of the cases $|\mathcal{H}_i^*(t)| < |\mathcal{H}_i(t)|$). Finally, the computational complexity for GTrack is the sum of $O(N^3)$ which is required by the Hungarian algorithm and $O(N^2)$ which is required to solve the graph, thus resulting in an overall upper limit of $O(N^3)$.

## 4.5 Summary

We proposed a detection and tracking algorithm for the localisation of interacting people. Compared to the state-of-the-art approach Šochman and Hogg [105] that performs group detection by only analysing the Social Force Model forces, we embedded in the algorithm two interaction constraints that modelled typical behaviours of interacting people. The first restricted the interaction among people to only those walking in the same direction and the second modelled the approaching of a person to a stationary group. The method analysed direction and velocity of people over time and groups were detected with a short latency of 0.4 seconds. In addition to this, the temporal consistency of the localised groups was improved with a graph-based tracker that linked the centres of interaction with a latency of 1 second. We showed that our framework outperformed state-of-the-art methods [10, 105, 118] on BIWI-ETH, BIWI-HOTEL and Student003 datasets that presented low- and mid-density crowds.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 4.8: Samples of group tracking results obtained with SFM-TR and GTrack on the (a)-(c) BIWI-ETH, (d)-(f) BIWI-HOTEL, and (g)-(i) Student003 datasets.

# Chapter 5

# Conclusions

## 5.1 Summary of the achievements

In this thesis, we addressed the problem of person re-identification and interaction localisation of people using human motion models with the aim of enhancing automatic video surveillance. Potential applications for our work could range from improving security in an existing environment to data analysis for scene understanding or for commercial purposes, as discussed below. Given a set of cameras with non-overlapping FOVs, the task of following a person across all cameras is very challenging, especially in those situations where people move in a crowd and look very similar to each other. For this reason, we proposed to employ human motion models to predict the paths followed by people in unobserved regions and to create candidate locations for reappearance. We exploited and modelled the fact that walking people are normally attracted by regions of interest (exits, shops, seats and meeting points) that are common to the majority, and we designed a parametric algorithm that does not need any data-driven learning, thus avoiding the need of a training set. The predictions generated can be used in the person re-identification task and provide a surveillance operator with the potential paths people follow when unobserved. These potential paths can be used for understanding movement in areas that cameras cannot monitor and to estimate the most likely movements of people. Unlike existing methods for crowd simulation where the simulation is performed by fixing a set of parameters beforehand, in our method the set of hypotheses for people's movements are generated using the velocity of people in the monitored scene, thus allowing estimations that are more realistic and, hence, closer to the actual people's movements. A possible application could be the improvement of safety in a building by, for instance, avoiding the placement of barriers in locations that obstruct the actual flow of people. The safety of a building is normally analysed before the building is constructed using crowd simulations and, after the building is constructed, using the video streams from the camera network; however our method could provide estimations from real data also in those areas that are not in the FOV of any camera.

In particular, for person re-identification, we critically analysed and reorganised the state of the art based on appearance features, cross-camera calibration and the approaches used for inter-camera association. We proposed and validated two person re-identification methods based on two models for people's movement in unobserved regions that exploited the map of the environment. The first model, Landmark-Based Model (LBM), was based on landmarks corresponding to interest and crossing regions in the site, and the second model, Multi-Goal Social Force Model (MG-SFM), was based on a modification of the Social Force Model that took into account barrier avoidance as well as the desire of people to move towards specific goals. We also presented a person's representation that is appropriate for crowded scenarios, and based on a vertical strip covering the upper body and containing the head. A challenging dataset from London Gatwick airport was employed for comparing the proposed methods with state-of-the-art approaches in a realistic and crowded scenario. Re-identification results of the motion models showed similar trends in the Cumulative Matching Characteristic (CMC) curves, even if LBM had lower complexity in terms of modelling and MG-SFM was more suited to modelling complex aspects of people's movement. Appearance-based methods that extracted features from the full body performed poorly compared to algorithms that extracted features from the upper body and to those that only considered possible variations on the travelling time of people in the unobserved regions (spatio-temporal modelling). Finally, the best performance for person re-identification was obtained by combining spatio-temporal and appearance cues.

In addition to this, motivated by the fact that 50-70% of walking activity of people takes place in groups, we localise interacting people in single camera views by analysing a human motion model that describes expected people's movements. Also in this model, no data-driven learning is necessary. On the one hand, the localisation of interacting people over time can be used to redirect the focus of attention of a surveillance operator to those areas where interactions are not supposed to happen. On the other hand, offline analysis of group formations can be used for multiple applications. For instance, when fights and similar dangerous interactions occur, an automatic analysis of the recordings would help in understanding where and when the event started, so the protagonists could be identified and similar situations could be avoided in the future. The understanding of where groups are more likely to form could also be used to infer the locations where people prefer to interact and socialise. This scene understanding allows the localisation of areas that should not be substantially modified in any future refurbishment because people already use and socialise in them, while if there are areas that are not used frequently this could indicate that they are in need of improvement. The localisation of these areas and the analysis of where people are more likely to move when in a group, is also very important for commercial and marketing purposes. For instance, advertisements can be effectively placed in those areas where groups of people move in order to maximise the number of people reached and to promote those products/activities specific for groups of people.

In particular, we proposed to embed two constraints for group formation in an interaction localisation algorithm based on Social Force Model. The Social Force Model described the forces involved in people's movement that are: the forces to maintain a desired velocity, to keep people at a comfortable distance from other people and barriers, and to keep people together when in a group. Interaction localisation was performed using an algorithm that iteratively analysed all the forces acting on people at a certain time instant and calculated the expected movement that people should have when interacting with other people. In order to improve group detection, we proposed to limit the interactions to those people moving in the same direction and to those decelerating when approaching a static group. After interactions were localised, a graph-based algorithm was applied to link over time the group centroid and to follow those interactions that were consistent over time. The improvements on group localisation of the proposed approach with respect to state-of-the-art methods were shown quantitatively using the one minus False Positive rate (1-FP) and the Group Detection Success Rate (GDSR) metrics, and qualitatively on three datasets presenting a different number of interactions.

In summary, we demonstrated that solutions for open problems related to the monitoring of human movements can be designed by understanding people's movement and modelling them without the need of data-driven learning. The applications of our motion prediction and interaction localisation of people in crowds could be video surveillance for re-identification and group detection, safety that improves the quality of an environment by understanding where people move when unobserved and when in a group, and commercial purposes that exploit data estimation and analysis to obtain the most likely paths people follow.

## 5.2 Future work

The future direction of our work are summarised below:

1. The application of the methods proposed in this thesis is limited by the fact that single-camera multi-person detection and tracking were considered solved, and manual annotations for people's trajectories were used in the experiments. In order to input automatically extracted trajectories to the proposed motion prediction and interaction localisation algorithms, their robustness to imprecise people's location, and false positive and false negative trajectories, needs to be evaluated [55, 79, 98]. Our proposed solution for single-camera detection and tracking is reported in App. A [J1].
2. Occlusion is one of the main issues that a single-camera tracking algorithm has to address [57]. The motion propagation algorithms presented in Ch. 3 could be used to provide location candidates for people's reappearance after a long occlusion. Unlike Gong *et al.* [34] that use a path planning method, a motion model designed for people, like the MG-SFM, may result in a more accurate motion estimation. Furthermore, similarly to Jin and Bhanu [49] grouping information could be integrated into the motion propagation framework.
3. The motion prediction models proposed in Ch. 3 could be applied in different scenarios only if interest and key regions, and goals are correctly localised in the specific site. The automatic definition of these regions in the observed areas [107] and, building on Idrees *et al.* [44] that estimate car behaviours in unobserved regions, the possible configurations of the unobserved areas could be integrated in the motion models.
4. The propagation error of the motion prediction methods could be analysed and the scalability of the methods evaluated using a larger camera network. Moreover, a dataset with people that only appear in one camera and do not reappear in other cameras, could be employed in order to analyse the robustness of the method when applied to this challenging situation [9].
5. Since in Ch. 3 only one set of appearance features was employed for person re-identification, further tests could evaluate how different appearance features affect the performance

of the re-identification based on motion prediction.

6. Similarly to Leal-Taixé *et al.* [59] and Pellegrini *et al.* [79], the interaction localisation method presented in Ch. 4 could be embedded in a target tracking framework. Tracking could benefit from a human motion model that better describes people's movement than linear motion. Moreover, the information of people walking in a group could be used to generate more robust cues for person re-identification in a multi-camera tracking framework [87, 120].

7. Since real applications may require the localisation of interacting people a few time instants in advance [111], we could perform interaction prediction by creating a set of hypotheses for future interactions among people. Past people's trajectories define actual groups and future interactions are estimated by propagating this information, thus permitting the understanding of which people will interact and, for instance, join an already formed group.

# Appendix A

# Multi-person tracking on confidence maps: an application to people tracking

## A.1 Introduction

Multi-target tracking is a challenging task in real scenarios due to the variability of target movements, shapes and sizes over time, clutter and occlusions. Moreover, the computational cost may exponentially increase with the number of co-occurring targets to be tracked and the maximum number of targets has to be fixed *a priori* [J1]. Compared to single-target tracking where the state of each target is represented by a single state vector [116], for multi-target tracking either the state vector is increased with respect to the number of targets [11, 14, 41, 56, 57, 63, 70, 91, 112, 113, 114] or a single-target tracking is initialised for each target [1, 17, 24, 52, 100, 104, 109, 115]. We refer to the two approaches as *one-state-per-target* (OSPT) and *one-filter-per-target* (OFPT) methods, respectively. OSPT methods perform the tracking optimisation at each time step on the overall state space. In this case, only a limited number of targets can be tracked due to a prior definition of the maximum number of allowed targets [24] or ad-hoc stages used to estimate the number of targets in the scene [14, 70]. OFPT methods perform tracking by a local optimisation for each target, thus limiting its application to situations where the number of targets is small and targets are easily distinguishable.

We propose a multi-target tracker based on track-before-detect algorithm [89] and applied to confidence maps (MT-TBD) [J1]. To allow for multi-target tracking, we develop a method where target IDs are assigned by using Mean-Shift clustering and Gaussian Mixture Model (GMM), and the birth and death of targets are modelled with a Markov Random Field (MRF). Unlike Buzzi *et al.* [19], we do not need to define the maximum number of targets *a priori* and, unlike Breitenstein *et al.* [17], the initialisation of a track may occur in any location of the image, thus making the MT-TBD completely automatic and flexible to different scenarios. MRF allows multi-target tracking without the augmentation of the state (OSPT methods, like the work by Boers and Driessen [14]) or the number of filters (OFPT methods), caused by an increase in the number of targets. Moreover, the use of MRF overcomes the limitations of Buzzi *et al.* [19] by allowing a reliable tracking of close targets without loss of performance and the formulation with a MRF leads to a computational complexity depending only on the number of particles. We apply the MT-TBD to people tracking using a postprocessing phase on a temporal window that employs track duration, background information and people's appearance. Compared to the recent work Benfold and Reid [11], the tracking accuracy improved by 11% with 2 seconds of latency and by 10% with 4 seconds of latency on a dataset from the town centre of Oxford, UK.

## A.2 Related work

In this section, we discuss recent works on multi-person tracking, we analyse their main contributions and classify each method in its corresponding category. Multi-target video trackers can be classified into causal and non-causal methods. *Causal* methods use information from past and present observations to estimate trajectories at the current time step. *Non-causal* methods use also information from future observations, thus resulting in a delayed decision. Although non-causal approaches are not suitable for time-critical applications, they can achieve a global optimum leading to more robust results during occlusions.

Causal trackers can be for example Bayesian filters [1, 11, 17, 109, 115]. Yang *et al.* [115] use a Bayesian detection association obtained by Convolutional Neural Network (CNN) trained on colour histograms, elliptical head model, and bags of SIFTs. Benfold and Reid [11] find the optimum trajectories within a four-second window by a Minimum Description Length (MDL) method applied on trajectories from a forward and backward Kanade-Lucas-Tomasi (KLT) tracking and from a Markov Chain Monte Carlo Data Association (MCMCDA). Alternatively, a particle filter is used in [1, 17, 109]. Ali and Dailey [1] track heads obtained by Haar-like features and AdaBoost; Xing *et al.* [109] employ the Hungarian algorithm for the optimisation of short but reliable trajectories obtained by tracking the upper human body. Breitenstein *et al.* [17] track, depending on the scenario, people detected by Histogram of Oriented Gradients (HOG) or Implicit Shape Model (ISM), where the association between detections and tracks is performed by a greedy algorithm and boosting. A different approach is presented in Rodriguez *et al.* [91] where tracking is obtained on four points per head by KLT and head detection is optimised by crowd density estimation and camera-scene geometry. Tag-and-track methods for a high-density crowd are proposed in [3, 90], where targets are assumed to follow a learned crowd behaviour. Ali and Shah [3] deal with crowds with coherent motion by modelling their global behaviour, the environment structure and the local behaviour of people. Rodriguez *et al.* [90] focus on crowds with non-coherent motion where the modelling is performed by Correlated Topic Model (CTM) that predicts the next position of a person by exploiting the optical flow. Note that among causal methods, only Benfold and Reid [11] and Rodriguez *et al.* [91] use an OSPT framework. This is because the OSPT is generally more complex than OFPT, but the modelling for multi-person tracking is more flexible and computationally cheaper [11].

Among non-causal trackers, short term tracks (tracklets) [41, 56, 57, 63, 112, 113, 114] can be associated over time by using a modification of the Multi-Hypothesis Tracking (MHT) algorithm [88] where the detections are obtained by the Wu *et al.* [108] person detector. Huang *et al.* [41] associate tracklets by Hungarian algorithm using position, time and appearance features, and then refine them using entry and exit points in the scenes, which are in turn learned from tracklets. Li *et al.* [63] show how the association can be improved by using a combination of RankBoost and AdaBoost in a hierarchical approach where, by starting from the lower levels, longer trajectories are generated using a set of 14 features per tracklet. In Yang *et al.* [112], the association is performed using RankBoost applied to an optimisation of affinities and dependencies between tracklets by a Conditional Random Field (CRF). Kuo *et al.* [56] associate tracklets using an AdaBoost classifier which learns online the discriminative appearance of targets based on colour histogram, covariance matrix features and HOG. Moreover, Kuo *et al.* [57] extract motion, time and appearance from different body parts of each target in order to perform a re-identification step to resolve long-term occlusions. Yang and Nevatia [113] learn online the non-linear motion of people and a Multiple Instance Learning (MIL) framework for the appearance modelling using the estimation of entry and exit regions. Furthermore, Yang and Nevatia [114] use CRF to model affinity relationships between pairs of tracklets, where the association of tracklets is based on Hungarian algorithm and a heuristic search.

Table A.1: Summary of recent state-of-the-art and proposed [J1] methods for multi-person tracking, and datasets used (see text for details). Legend: CM = Confidence Map; OSPT = One-State-Per-Target; CRF = Conditional Random Field; OLDAMs = Online Learning of Discriminative Appearance Models; PIRMPT = Person Identity Recognition based Multi-Person Tracking; MIL = Multiple Instance Learning; KLT = Kanade-Lucas-Tomasi feature tracker; MCMCDA = Markov-Chain Monte-Carlo Data Association; JPDA = Joint Probabilistic Data Association; iLids = i-LIDS dataset from Westminster subway station (London, UK); TRECVID = i-LIDS dataset from London Gatwick airport.

| **Ref.** | **Method** | **CM** | **OSPT** | **Causality** | **Dataset** |
|---|---|---|---|---|---|
| [41] | Three-stage algorithm, Hungarian algorithm | | ✓ | | CAVIAR, iLids |
| [56] | AdaBoost on OLDAMs | | ✓ | | CAVIAR, TRECVID |
| [57] | PIRMPT | | ✓ | | CAVIAR, ETH, TRECVID |
| [63] | HybridBoost | | ✓ | | CAVIAR, TRECVID |
| [112] | CRF, RankBoost | | ✓ | | TRECVID |
| [113] | Learning of motion map, MIL for appearance | | ✓ | | CAVIAR, PETS2009, TRECVID |
| [114] | CRF, Hungarian algorithm/heuristic search | | ✓ | | ETH, TRECVID, TUD |
| [11] | KLT, MCMCDA | ✓ | ✓ | ✓ | iLids, PETS2007, TownCentre |
| [58] | Automatic relevance detection, JPDA | ✓ | ✓ | ✓ | Ants, laser output |
| [91] | KLT points, Crowd density estimation | ✓ | ✓ | ✓ | Political rally |
| [17] | Particle filter, Greedy algorithm, Boosting | ✓ | | ✓ | iLids, PETS2009, soccer, TUD campus, UBC Hockey |
| [1] | Particle filter | | | ✓ | Bangkok station |
| [3] | Floor fields | | | ✓ | Marathon, train station |
| [90] | Correlated Topic Model | | | ✓ | Mall, sport crowd |
| [109] | Particle filter, Hungarian algorithm | | | ✓ | CAVIAR, ETH |
| [115] | Bayesian filter, Hungarian algorithm | | | ✓ | CAVIAR, TRECVID |
| [J1] | Multi-target track-before-detect | ✓ | | ✓ | APIDIS, ETH, iLids, TownCentre, TRECVID |

Our proposed MT-TBD, similarly to Stalder *et al.* [96] and Breitenstein *et al.* [17], is a causal method that makes use of confidence maps as a measurement for tracking. However, compared to Stalder *et al.* [96], we use the confidence maps online without the need of any temporal processing and, compared to Breitenstein *et al.* [17], an automatic assignment between the confidence map and targets is performed. In addition, unlike Breitenstein *et al.* [17] which uses manually selected areas at the borders of the image to initialise tracks, we do not use any prior information about the scene. This becomes extremely advantageous when targets temporarily undergo a total occlusion in any position of the image. We overcome the limitations of OFPT approaches [17, 58] with a global and instantaneous optimisation of target tracking in the MT-TBD by employing a general likelihood function obtained from a controlled sequence. Finally, unlike De Leat *et al.* [58], the use of multiple measurements per target is tested in various crowded scenes with different camera perspectives.

Table A.1 summarises the methods covered in this section and the dataset on which these methods have been tested.

## A.3 Results and analysis

We show the results of the MT-TBD as multi-person single-camera tracking on automatically generated confidence maps. We use the TownCentre dataset[1] composed of 4500 frames of size 1980 × 1080 pixels, recorded from Oxford (UK) town centre at 25 Hz. For a fair comparison with Benfold and Reid [11], we use the head locations provided by the authors, which are generated using HOG features and SVM. As the provided person's locations have already been thresholded, they are not in the form of intensity levels. For this reason, the input to the MT-TBD is a confidence map with 2D Deltas in correspondence to each localised head.

Given a bounding box for each target along with the 2D Deltas at each time step, a true positive track is defined as the one having a bounding box overlapping at least 25% with the ground truth [11]. Let $tp$ be the number of all the true positive tracks in a video sequence, $fp$ all the false positive tracks, $fn$ all the false negative tracks, $IDS$ the number of all ID switches, and $N_G$ the number of ground truth targets. Performance evaluation is obtained by calculating the Multiple Object Tracking Accuracy (MOTA) and the Multiple Object Tracking Precision (MOTP), Precision and Recall [13]. MOTA is calculated as

$$MOTA = 1 - \frac{(N_G - tp) + fp + IDS}{N_G} \tag{A.1}$$

and MOTP as

$$MOTP = \frac{O_t}{N_m}, \tag{A.2}$$

where $O_t$ quantifies the overlap between the tracked bounding boxes at each time instant $t$ and the ground-truth bounding boxes, and $N_m$ is the number of ground-truth targets mapped with the tracking output for the whole video sequence. Precision is calculated as

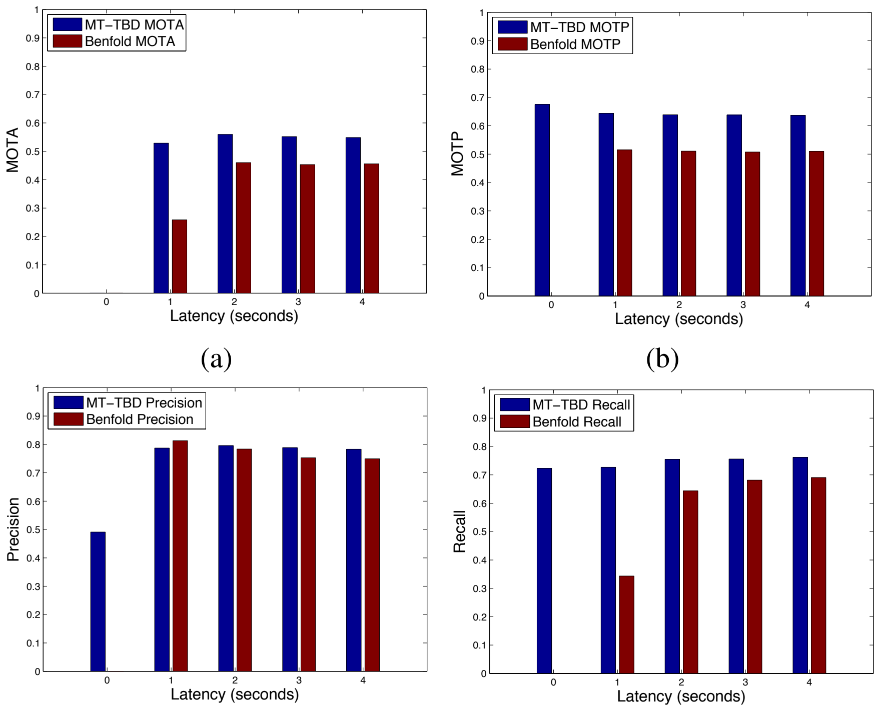
$$P = \frac{tp}{tp + fp} \tag{A.3}$$

and Recall as

$$R = \frac{tp}{tp + fn}. \tag{A.4}$$

We show how our method outperforms the recent work Benfold and Reid [11] by using

[1] http://www.robots.ox.ac.uk/ActiveVision/Research/Projects/2009bbenfold_headpose/project.html. Last accessed: March 2012.



(a) (b)

(c) (d)

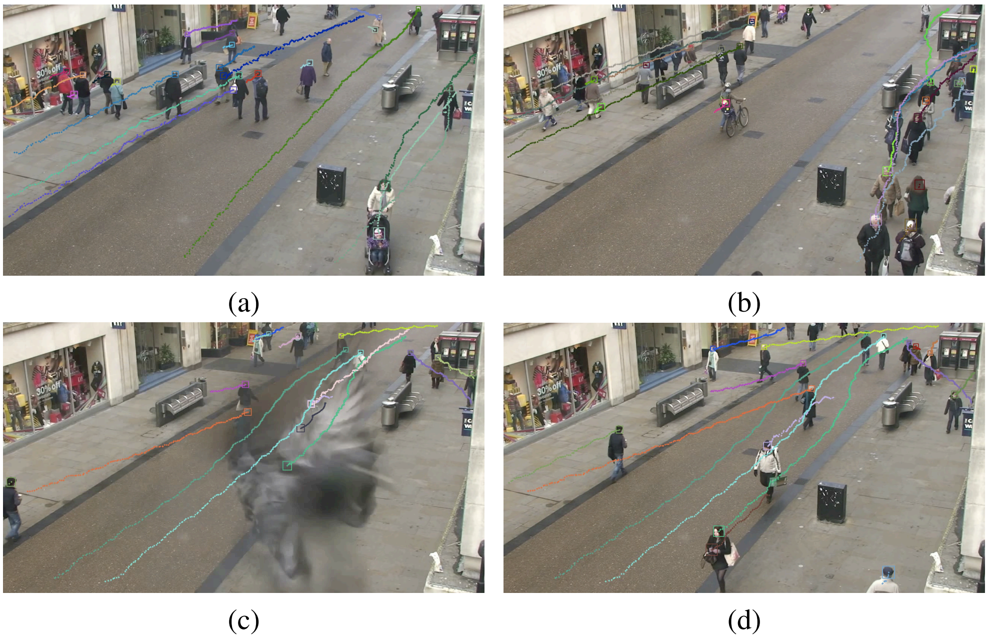
Figure A.1: Comparison of the results on TownCentre dataset with Benfold and Reid method [11]. The graphs show the variation of the scores as a function of the latency introduced by the postprocessing: (a) MOTA, (b) MOTP, (c) Precision and (d) Recall.

the same observations for tracking. This scenario is fairly challenging as it contains very close targets and the FOV of the camera is very large, hence ID switches are likely to be frequent. For comparison, we present the results with the same latency used in Benfold and Reid [11] for postprocessing and, in particular, of 1, 2, 3, and 4 seconds (1 second = 25 frames). In order to show the global improvement of our proposed method, we also include the performance of the MT-TBD without any postprocessing. Note that, unlike our tracker, the work in Benfold and Reid [11] cannot work with latency equal to 0.

Figure A.1 shows the quantitative results. The superior performance of the proposed method is highlighted by the value of Recall that is consistently higher than Benfold and Reid [11] at various latencies. For the MT-TBD without latency (and no postprocessing), the value of Recall is already high and comparable with 4 seconds of latency. However, the Precision in this case is lower due to the short and false tracks generated by the temporally-consistent false positive head locations. By applying the proposed postprocessing, the Precision drastically increases. Table A.2 summarises the final results and Fig. A.2 shows sample tracking results, where it is clear that the method is robust under severe occlusions with a few fragmented tracks.

Table A.2: Comparison between the results obtained using the proposed method (MT-TBD) and Benfold2011 [11] on TownCentre dataset [11]. The number of frames between round brackets represents the temporal window duration used for postprocessing. Key - IDS: ID Switches.

| **Method** | **MOTA** | **MOTP** | **Precision** | **Recall** | **IDS** |
|---|---|---|---|---|---|
| MT-TBD (100frs) | 0.546 | 0.637 | 0.783 | 0.762 | 285 |
| Benfold2011 | 0.454 | 0.508 | 0.738 | 0.710 | - |



(a)

(b)

(c)

(d)

Figure A.2: Sample tracking results of the proposed method on TownCentre dataset [11]. The tracks are shown from the initialisation of the track.

## A.4 Summary

In this appendix, we described a Bayesian method for multi-object tracking based on *track-before-detect*, which utilises a Markov Random Field applied on the particles to perform tracking *(i)* of unknown and large number of targets, and *(ii)* by probabilistically managing the ID assignment to avoid ID switches with close targets. The state estimate of a target is performed via Mean-Shift clustering and supported by Mixture of Gaussians in order to enable an accurate assignment of IDs within each single cluster. The birth and death of the targets at each iteration of the filter is modelled with a Markov Random Field. The robustness of our algorithm was demonstrated by applying the method on a surveillance dataset obtaining better results with respect to a recent method from the state-of-the-art.

# Bibliography

[1] I. Ali and M. N. Dailey. Multiple human tracking in high-density crowds. In *Proc. of Conference on Advanced Concepts for Intelligent Vision Systems*, pages 540–549, Bordeaux, France, 28 September-2 October 2009.

[2] S. Ali and M. Shah. A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, Minneapolis, MN, USA, 18-23 June 2007.

[3] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *Proc. of European Conference on Computer Vision*, pages 1–14, Marseille, France, 12-18 October 2008.

[4] E. L. Andrade and R. B. Fisher. Simulation of crowd problems for computer vision. In *First Int. Workshop on Crowd Simulation*, pages 71–80, Lausanne, Switzerland, 24-25 November 2005.

[5] G. Antonini, S. V. Martinez, M. Bierlaire, and J. P. Thiran. Behavioral priors for detection and tracking of pedestrians in video sequences. *International Journal of Computer Vision*, 96(2):159–180, 2006.

[6] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person re-identification using haar-based and DCD-based signature. In *Workshop on Activity Monitoring by Multi-Camera Surveillance Systems in conjunction with IEEE Int. Conference on Advanced Video and Signal Based Surveillance*, pages 1–8, Boston, MA, USA, 29 August-1 September 2010.

[7] D. Bauer, S. Seer, and N. Brändle. Macroscopic pedestrian flow simulation for designing crowd control measures in public transport after special events. In *Summer Computer Simulation Conference*, pages 1035–1042, San Diego, CA, USA, 15-18 July 2007.

[8] M. Bäuml, K. Bernardin, M. Fischer, and H. K. Ekenel. Multi-pose face recognition for person retrieval in camera networks. In *IEEE Int. Conference on Advanced Video and Signal Based Surveillance*, pages 441–447, Boston, Massachusetts, 29 August-1 September 2010.

[9] M. Bäuml and R. Stiefelhagen. Evaluation of local features for person re-identification in image sequences. In *IEEE Int. Conference on Advanced Video and Signal Based Surveillance*, pages 291–296, Klagenfurt, Austria, 30 August-2 September 2011.

[10] L. Bazzani, M. Cristani, and V. Murino. Decentralized particle filter for joint individual-group tracking. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1886–1893, Providence, RI, USA, 16-21 June 2012.

[11] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3457–3464, Colorado Springs, USA, 20-25 June 2011.

[12] G. Berdugo, O. Soceanu, Y. Moshe, D. Rudoy, and I. Dvir. Object reidentification in real world scenarios across multiple non-overlapping cameras. In *European Signal Processing Conference*, Aalborg, Denmark, 23-27 August 2010.

[13] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008(1):246–309, February 2008.

[14] Y. Boers and J. N. Driessen. Multitarget particle filter track-before-detect applications. *IEE Proc. Radar, Sonar and Navigation*, 151(6):351–357, December 2004.

[15] P. V. K. Borges, N. Conci, and A. Cavallaro. Video-based human behavior understanding: a survey. *IEEE Trans. on Circuits and Systems for Video Technology*, DOI: `http://dx.doi.org/10.1109/TCSVT.2013.2270402`, available online 20 June 2013.

[16] R. Bowden and P. KaewTraKulPong. Towards automated wide area visual surveillance: tracking objects between spatially-separated, uncalibrated views. *IEE Proc. on Vision, Image and Signal Processing*, 152(2):213–223, April 2005.

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

[18] G. J. Brostow and R. Cipolla. Unsupervised Bayesian detection of independent motion in crowds. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 594–601, New York, NY, USA, 17-22 June 2006.

[19] S. Buzzi, M. Lops, L. Venturino, and M. Ferri. Track-before-detect procedures in a multi-

target environment. *IEEE Trans. of Aerospace and Electronic Systems*, 44(3):1135–1150, July 2008.

[20] M.-C. Chang, N. Krahnstoever, and W. Ge. Probabilistic group-level motion analysis and scenario recognition. In *Proc. of IEEE International Conference on Computer Vision*, pages 747–754, Barcelona, Spain, 6-13 November 2011.

[21] K.-W. Chen, C.-C. Lai, P.-J. Lee, C.-S. Chen, and Y.-P. Hung. Adaptive learning for target tracking and true linking discovering across multiple non-overlapping cameras. *IEEE Trans. on Multimedia*, 13(4):625–638, August 2011.

[22] Y. Cheng, W. Zhou, Y. Wang, C. Zhao, and S. Zhang. Multi-camera-based object handoff using decision-level fusion. In *Int. congress on Image and Signal Processing*, pages 1–5, Tianjin, China, 17-19 October 2009.

[23] A. Colombo, J. Orwell, and S. Velastin. Colour constancy techniques for re-recognition of pedestrians from multiple surveillance cameras. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications In conjunction with the European Conference on Computer Vision*, Marseille, France, 18 October 2008.

[24] J. Czyz, B. Ristic, and B. Macqa. A particle filter for joint detection and tracking of color objects. *Image and Vision Computing*, 25(8):1271–1281, August 2007.

[25] CAVIAR dataset. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/. Last accessed: March 2012.

[26] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing*, 2(2):127–151, June 2011.

[27] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(12):2179–2195, December 2009.

[28] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2360–2367, San Francisco, USA, 13-18 June 2010.

[29] G. A. Frank and C. O. Dorso. Room evacuation in the presence of an obstacle. *Physica A: Statistical Mechanics and its Applications*, 390(11):2135–3145, June 2011.

[30] W. Ge, R. T. Collins, and B. Ruback. Automatically detecting the small group structure of a crowd. In *IEEE Workshop on Applications of Computer Vision*, pages 1–8, Snowbird, UT, USA, 7-8 December 2009.

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

[32] N. Gheissari, T. B. Sebastian, P. H. Tu, J. Rittscher, and R. Hartley. Person reidentification using spatiotemporal appearance. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1528–1535, New York, NY, USA, 17-22 June 2006.

[33] A. Gilbert and R. Bowden. Incremental, scalable tracking of objects inter camera. *Computer Vision and Image Understanding*, 111(1):43–58, July 2008.

[34] H. Gong, J. Sim, M. Likhachev, and J. Shi. Multi-hypothesis motion planning for visual object tracking. In *Proc. of IEEE International Conference on Computer Vision*, pages 619–626, Barcelona, Spain, 6-13 November 2011.

[35] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. of European Conference on Computer Vision*, pages 262–275, Marseille, France, 12-18 October 2008.

[36] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *ACM/IEEE Int. Conference on Distributed Smart Cameras*, pages 1–6, California, USA, 7-11 September 2008.

[37] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Second ed. Cambridge University Press (UK), 2004.

[38] D. Helbing, I. Farkas, and T. Vicsek. Simulating dynamical features of escape panic. *Nature*, 407:487–490, September 2000.

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

[40] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *Proc. of European Conference on Computer Vision*, pages 780–793, Florence, Italy, 7-13 October 2012.

[41] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *Proc. of European Conference on Computer Vision*, pages 788–801, Marseille, France, 12-18 October 2008.

[42] R. L. Hughes. The flow of human crowds. *Annual Review of Fluid Mechanics*, 35:169–182, 2003.

[43] i-LIDS. *Home Office multiple-camera tracking scenario definition (UK).* 2008.

[44] H. Idrees, I. Saleemi, and M. Shah. Statistical inference of motion in the invisible. In *Proc. of European Conference on Computer Vision*, pages 544–557, Florence, Italy, 7-13 October 2012.

[45] O. Javed, Z. Rasheed, O. Alatas, and M. Shah. Knight$^M$: A real time surveillance system for multiple overlapping and non-overlapping cameras. In *IEEE Conference on Multimedia and Expo*, pages I.649–I.652, Baltimore, MD, USA, 6-9 July 2003.

[46] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. Tracking across multiple cameras with disjoint views. In *Proc. of IEEE International Conference on Computer Vision*, pages 952–957, Nice, France, 14-17 October 2003.

[47] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162, February 2008.

[48] K. Jeong and C. Jaynes. Object matching in disjoint cameras using a colour transfer approach. *Springer Journal of Machine Vision and Applications*, 19(5):88–96, September 2008.

[49] Z. Jin and B. Bhanu. Integrating crowd simulation for pedestrian tracking in a multi-camera system. In *Int. Conference on Distributed Smart Cameras*, pages 1–6, Hong Kong, 30 October-2 November 2012.

[50] A. Johansson, D. Helbing, and P. K. Shukla. Specification of a microscopic pedestrian model by evolutionary adjustment to video tracking data. *Advances in Complex Systems*, 10(2):271–288, December 2007.

[51] V. Kettnaker and R. Zabih. Bayesian multi-camera surveillance. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 252–259, Fort Collins, CO, USA, 23-25 June 1999.

[52] Z. Khan, T. Balch, and F. Dellaert. MCMC data association and sparse factorization updating for real time multitarget tracking with merged and multiple measurements. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(12):1960–1972, December 2006.

[53] K. M. Kitani, D. Bagnell, and M. Hebert. Activity forecasting. In *Proc. of European Conference on Computer Vision*, pages 201–214, Firenze, Italy, 7-13 October 2012.

[54] T. Ko. A survey on behavior analysis in video surveillance for homeland security applications. In *IEEE Applied Imagery Pattern Recognition Workshop*, pages 1–8, Washington, DC, USA, 15-17 October 2008.

[55] C.-H. Kuo, C. Huang, and R. Nevatia. Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In *Proc. of European Conference on Computer Vision*, pages 383–396, Hersonissos, Crete, Greece, 5-11 September 2010.

[56] C.-H. Kuo, C. Huang, and R. Nevatia. Multi-target tracking by on-line learned discriminative appearance models. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 685–692, San Francisco, CA, USA, 13-18 June 2010.

[57] C.-H. Kuo and R. Nevatia. How does person identity recognition help multi-person tracking? In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1217–1224, Colorado Springs, CO, USA, 20-25 June 2011.

[58] T. De Laet, H. Bruyninckx, and J. De Schutter. Shape-based online multitarget tracking and detection for targets causing multiple measurements: Variational bayesian clustering and lossless data association. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(12):2477–2491, December 2011.

[59] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *Proc. of IEEE International Conference on Computer Vision Workshops*, pages 120–127, Barcelona, Spain, 6-13 November 2011.

[60] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. *Computer Graphics Forum (Proc. of Eurographics)*, 26(3):655–664, September 2007.

[61] V. Leung, J. Orwell, and S. A. Velastin. Performance evaluation of re-acquisition methods for public transport surveillance. In *Int. Conference on Control, Automation, Robotics and Vision*, pages 705–712, Hanoi, Vietnam, 17-20 December 2008.

[62] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *Proc. of Asian Conference on Computer Vision*, pages 31–44, Daejeon, South Korea, 5-9 November 2012.

[63] Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2953–2960, Miami, FL, USA, 20-25 June 2009.

[64] G. Lian, J. Lai, C. Y. Suen, and P. Chen. Matching of tracked pedestrians across disjoint camera views using CI-DLBP. *IEEE Trans. on Circuits and Systems for Video Technology*, 22(7):1087–1099, July 2012.

[65] C. C. Loy, T. Xiang, and S. Gong. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90(1):106–129, October 2010.

[66] M. Luber, J. A. Stork, G. D. Tipaldi, and K. O. Arras. People tracking with human motion predictions from social forces. In *Int. Conference on Robotics and Automation*, pages 464–469, Anchorage, AK, USA, 3-8 May 2010.

[67] M. Luber, G. D. Tipaldi, and K. O. Arras. Place-dependent people tracking. *The Int. Journal of Robotics Research*, 30(30):280–293, March 2011.

[68] F. Lv, T. Zhao, and R. Nevatia. Camera calibration from video of a walking human. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(9):1513–1518, September 2006.

[69] C. Madden, E. D. Cheng, and M. Piccardi. Tracking people across disjoint camera views by an illumination tolerant appearance representation. *Springer Journal of Machine Vision and Applications*, 18(3):233–247, May 2007.

[70] E. Maggio and A. Cavallaro. Learning scene context for multiple object tracking. *IEEE Trans. on Image Processing*, 18(8):1873–1884, August 2009.

[71] D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages II.205–II.210, Washington, DC, USA, 27 June-2 July 2004.

[72] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942, Miami Beach, FL, USA, 20-25 June 2009.

[73] A. Millonig and K. Schechtner. Developing landmark-based pedestrian-navigation systems. *IEEE Trans. on Intelligent Transportation Systems*, 8(1):43–49, March 2007.

[74] M. Moussaïd, D. Helbing, S. Garnier, A. Johansson, M. Combe, and G. Theraulaz. Experimental study of the behavioural mechanisms underlying self-organization in human crowds. *Proc. of the Royal Society*, 276(1668):2755–2762, 7 August 2009.

[75] M. Moussaïd, N.Perozo, S. Garnier, D. Helbing, and G. Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PLoS ONE*, 5(4):e10047, 7 April 2010.

[76] I. Oliveira and J. Luiz. People re-identification in a camera network. In *IEEE Int. Conference on Dependable, Autonomic and Secure Computing*, pages 461–466, Chengdu, China, 12-14 December 2009.

[77] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010.

[78] S. Pellegrini. *Modeling and tracking social walkers*. PhD thesis, Department of Information Technology and Electrical Engineering, ETH Zurich, 2012.

[79] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *Proc. of European Conference on Computer Vision*, pages 452–465, Hersonissos, Crete, Greece, 5-11 September 2010.

[80] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proc. of IEEE International Conference on Computer Vision*, pages 261–268, Kyoto, Japan, 29 September-2 October 2009.

[81] S. Pellegrini, A. Ess, M. Tanaskovic, and L. Van Gool. Wrong turn - no dead end: A stochastic pedestrian motion model. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 15–22, San Francisco, CA, USA, 13-18 June 2010.

[82] F. Poiesi and A. Cavallaro. Detection and tracking of interacting targets. *IEEE Trans. on Image Processing*, (submitted), 2013.

[83] F. Porikli. Inter-camera color calibration by cross-correlation model function. In *Proc. of IEEE International Conference on Image Processing*, pages II.133–II.136 vol.3, Barcelona, Spain, 14-17 September 2003.

[84] B. Prosser, S. Gong, and T. Xiang. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *Proc. of British Machine Vision Conference*, pages 64.1–64.10, Leeds, UK, 1-4 September 2008.

[85] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *Proc. of British Machine Vision Conference*, pages 21.1–21.11, Aberystwyth, UK, 31 August-3 September 2010.

[86] Z. Qin and C. R. Shelton. Improving multi-target tracking via social grouping. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1972–1978, Providence, RI, USA, 16-21 June 2012.

[87] Z. Qin, C. R. Shelton, and L. Chai. Social grouping for target handover in multi-view video. In *Proc. of IEEE Int. Conference on Multimedia and Expo*, San Jose, CA, USA, 15-19 July 2013.

[88] D. Reid. An algorithm for tracking multiple targets. *IEEE Trans. on Automatic Control*, 24(6):843–854, December 1979.

[89] B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman filter: particle filters for tracking applications*. Artech House, Boston, 2004.

[90] M. Rodriguez, S. Ali, and T. Kanade. Tracking in unstructured crowded scenes. In *Proc. of IEEE International Conference on Computer Vision*, pages 1389–1396, Kyoto, Japan, 29 September-4 October 2009.

[91] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert. Density-aware person detection and tracking in crowds. In *Proc. of IEEE International Conference on Computer Vision*, pages 2423–2430, Barcelona, Spain, 6-13 November 2011.

[92] R. Satta, G. Fumera, F. Roli, M. Cristani, and V. Murino. A multiple component matching framework for person re-identification. In *Int. Conference on Image Analysis and Processing*, pages 140–149, Ravenna, Italy, 14-16 September 2011.

[93] P. Scovanner and M. F. Tappen. Learning pedestrian dynamics from the real world. In *Proc. of IEEE International Conference on Computer Vision*, pages 381–388, Kyoto, Japan, 29 September-2 October 2009.

[94] C. Siebler, K. Bernardin, and R. Stiefelhagen. Adaptive color transformation for person

re-identification in camera networks. In *ACM/IEEE Int. Conference on Distributed Smart Cameras*, pages 199–205, Hong Kong, 30 October-2 November 2010.

[95] B. Solmaz, B.E. Moore, and M. Shah. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(10):2064–2070, October 2012.

[96] S. Stalder, H. Grabner, and L. Van Gool. Cascaded confidence filtering for improved tracking-by-detection. In *Proc. of European Conference on Computer Vision*, pages 369–382, Hersonissos, Crete, Greece, 5-11 September 2010.

[97] G. K. Still. *Crowd Dynamics*. PhD thesis, University of Warwick, 2000.

[98] L. F. Teixeira, P. Carvalho, J. S. Cardoso, and L. Corte-Real. Automatic description of object appearances in a wide-area surveillance scenario. In *Proc. of IEEE International Conference on Image Processing*, pages 1609–1612, Orlando, FL, USA, 30 September-3 October 2012.

[99] L. F. Teixeira and L. Corte-Real. Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recognition Letters*, 320(2):157–167, January 2009.

[100] H. Tong, H. Zhang, H. Meng, and X. Wang. Multitarget tracking before detection via probability hypothesis density filter. In *Int. Conference on Electrical and Control Engineering*, pages 1332–1335, Wuhan, China, 25-27 June 2010.

[101] A. Turner and A. Penn. Encoding natural movement as an agent-based system: an investigation into human pedestrian behaviour in the built environment. *Environment and Planning B: Planning and Design*, 29(4):473–490, 2002.

[102] D. Vasquez, T. Fraichard, and C. Laugier. Incremental learning of statistical motion patterns with growing hidden markov models. *IEEE Trans. on Intelligent Transportation System*, 10(3):403–416, September 2009.

[103] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 511–518, Kauai, HI, USA, 8-14 December 2001.

[104] B.-N. Vo, B.-T. Vo, N.-T. Pham, and D. Suter. Joint detection and estimation of multiple

objects from image observations. *IEEE Trans. on Signal Processing*, 58(10):5129–5241, October 2010.

[105] J. Šochman and D. C. Hogg. Who knows who - inverting the social force model for finding groups. In *Proc. of IEEE International Conference on Computer Vision Workshop*, pages 830–837, Barcelona, Spain, 6-13 November 2011.

[106] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *Proc. of IEEE International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brasil, 14-20 October 2007.

[107] X. Wang, K. T. Ma, G.-W. Ng, and W. E. L. Grimson. Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. *International Journal of Computer Vision*, 95(3):287–312, December 2011.

[108] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, November 2007.

[109] J. Xing, H. Ai, and S. Lao. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1200–1207, Miami, FL, USA, 20-25 June 2009.

[110] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg. Who are you with and where are you going? In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1345–1352, Colorado Springs, CO, USA, 20-25 June 2011.

[111] X. Yan, I. Kakadiaris, and S. Shah. Predicting social interactions for visual tracking. In *Proc. of British Machine Vision Conference*, pages 102.1–102.11, Dundee, UK, 29 August-2 September 2011.

[112] B. Yang, C. Huang, and R. Nevatia. Learning affinities and dependencies for multi-target tracking using a CRF model. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1233–1240, Colorado Springs, CO, USA, 20-25 June 2011.

[113] B. Yang and R. Nevatia. Multi-target tracking by online learning of non-linear motion patterns and robust appearance model. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1918–1925, Providence, RI, USA, 16-21 June 2012.

[114] B. Yang and R. Nevatia. An online learned CRF model for multi-target tracking. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2034–2041, Providence, RI, USA, 16-21 June 2012.

[115] M. Yang, F. Lv, W. Xu, and Y. Gong. Detection driven adaptive multi-cue integration for multiple human tracking. In *Proc. of IEEE International Conference on Computer Vision*, pages 1554–1561, Kyoto, Japan, 29 September-2 October 2009.

[116] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(13):1–45, December 2006.

[117] Z. Yücel, F. Zanlungo, T. Ikeda, T. Miyashita, and N. Hagita. Deciphering the crowd: Modeling and identification of pedestrian group motion. *Sensors*, 13(1):875–897, January 2013.

[118] M. Zanotto, L. Bazzani, M. Cristani, and V. Murino. Online Bayesian nonparametrics for group detection. In *Proc. of British Machine Vision Conference*, Guilford, UK, 3-7 September 2012.

[119] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu. Crowd analysis: a survey. *Machine Vision and Applications*, 19(5-6):345–357, September 2008.

[120] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *Proc. of British Machine Vision Conference*, pages 13.1–13.11, London, UK, 7-10 September 2009.

[121] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(3):653–668, March 2013.

[122] B. Zhou, X. Wang, and X. M. Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2871–2878, Providence, RI, USA, 16-21 June 2012.