Loris Bazzani

# Beyond Multi-target Tracking

## Statistical Pattern Analysis of People and Groups

May 3, 2012

Università degli Studi di Verona
Dipartimento di Informatica

Advisor:
Prof. Vittorio Murino
Co-advisor:
Dr. Marco Cristani

Series N°: **TD-02-12**

Università di Verona
Dipartimento di Informatica
Strada le Grazie 15, 37134 Verona
Italy

*to my parents Lucia and Alberto,*
*my brother Johnny and*
*my fiancé Annarita*
*that supported me during these years.*

# Abstract

Every day millions and millions of surveillance cameras monitor the world, recording and collecting huge amount of data. The collected data can be extremely useful: from the behavior analysis to prevent unpleasant events, to the analysis of the traffic. However, these valuable data is seldom used, because of the amount of information that the human operator has to manually attend and examine. It would be like looking for a needle in the haystack.

The automatic analysis of data is becoming mandatory for extracting summarized high-level information (*e.g.*, John, Sam and Anne are walking together in group at the playground near the station) from the available redundant low-level data (*e.g.*, an image sequence). The main goal of this thesis is to propose solutions and automatic algorithms that perform high-level analysis of a camera-monitored environment. In this way, the data are summarized in a high-level representation for a better understanding. In particular, this work is focused on the analysis of moving people and their collective behaviors.

The title of the thesis, *beyond multi-target tracking*, mirrors the purpose of the work: we will propose methods that have the *target tracking* as common denominator, and go beyond the standard techniques in order to provide a high-level description of the data. First, we investigate the target tracking problem as it is the basis of all the next work. Target tracking estimates the position of each target in the image and its trajectory over time. We analyze the problem from two complementary perspectives: 1) the engineering point of view, where we deal with problem in order to obtain the best results in terms of accuracy and performance. 2) The neuroscience point of view, where we propose an attentional model for tracking and recognition of objects and people, motivated by theories of the human perceptual system.

Second, target tracking is extended to the camera network case, where the goal is to keep a unique identifier for each person in the whole network, *i.e.*, to perform *person re-identification*. The goal is to recognize individuals in diverse locations over different non-overlapping camera views or also the same camera, considering a large set of candidates. In this context, we propose a pipeline and appearance-based descriptors that enable us to define in a proper way the problem and to reach the-state-of-the-art results.

Finally, the higher level of description investigated in this thesis is the analysis (discovery and tracking) of social interaction between people. In particular, we focus on finding *small groups* of people. We introduce methods that embed notions of social psychology into computer vision algorithms. Then, we extend the detection of social interaction over time, proposing novel probabilistic models that deal with (joint) individual-group tracking.

# Sommario

Ogni giorno milioni e milioni di videocamere monitorano la vita quotidiana delle persone, registrando e collezionando una grande quantità di dati. Questi dati possono essere molto utili per scopi di video-sorveglianza: dalla rilevazione di comportamenti anomali all'analisi del traffico urbano nelle strade. Tuttavia i dati collezionati vengono usati raramente, in quanto non è pensabile che un operatore umano riesca a esaminare manualmente e prestare attenzione a una tale quantità di dati simultaneamente.

Per questo motivo, negli ultimi anni si è verificato un incremento della richiesta di strumenti per l'analisi automatica di dati acquisiti da sistemi di video-sorveglianza in modo da estrarre informazione di più alto livello (per esempio, John, Sam e Anne stanno camminando in gruppo al parco giochi vicino alla stazione) a partire dai dati a disposizione che sono solitamente a basso livello e ridondati (per esempio, una sequenza di immagini). L'obiettivo principale di questa tesi è quello di proporre soluzioni e algoritmi automatici che permettono di estrarre informazione ad alto livello da una zona di interesse che viene monitorata da telecamere. Così i dati sono rappresentati in modo da essere facilmente interpretabili e analizzabili da qualsiasi persona. In particolare, questo lavoro è focalizzato sull'analisi di persone e i loro comportamenti sociali collettivi.

Il titolo della tesi, *beyond multi-target tracking*, evidenzia lo scopo del lavoro: tutti i metodi proposti in questa tesi che si andranno ad analizzare hanno come comune denominatore il *target tracking*. Inoltre andremo oltre le tecniche standard per arrivare a una rappresentazione del dato a più alto livello. Per prima cosa, analizzeremo il problema del target tracking in quanto è alle basi di questo lavoro. In pratica, target tracking significa stimare la posizione di ogni oggetto di interesse in un immagine e la sua traiettoria nel tempo. Analizzeremo il problema da due prospettive complementari: 1) il punto di vista ingegneristico, dove l'obiettivo è quello di creare algoritmi che ottengono i risultati migliori per il problema in esame. 2) Il punto di vista della neuroscienza: motivati dalle teorie che cercano di spiegare il funzionamento del sistema percettivo umano, proporremo in modello attenzionale per tracking e il riconoscimento di oggetti e persone.

Il secondo problema che andremo a esplorare sarà l'estensione del tracking alla situazione dove più telecamere sono disponibili. L'obiettivo è quello di mantenere un identificatore univoco per ogni persona nell'intera rete di telecamere. In altre

parole, si vuole riconoscere gli individui che vengono monitorati in posizioni e telecamere diverse considerando un database di candidati. Tale problema è chiamato in letteratura re-indetificazione di persone. In questa tesi, proporremo un modello standard di come affrontare il problema. In questo modello, presenteremo dei nuovi descrittori di aspetto degli individui, in quanto giocano un ruolo importante allo scopo di ottenere i risultati migliori.

Infine raggiungeremo il livello più alto di rappresentazione dei dati che viene affrontato in questa tesi, che è l'analisi di interazioni sociali tra persone. In particolare, ci focalizzeremo in un tipo specifico di interazione: il raggruppamento di persone. Proporremo dei metodi di visione computazionale che sfruttano nozioni di psicologia sociale per rilevare gruppi di persone. Inoltre, analizzeremo due modelli probabilistici che affrontano il problema di tracking (congiunto) di gruppi e individui.

# Preface

This thesis summarizes the work I did during the three years and half of my PhD in collaboration with many other colleagues and friends. First, I would like to thank some special persons that without their contribution I will not make this goal possible. Then, in this section I will provide the list of the publications I have been involved in during my PhD. Some of the papers will also be discussed in this thesis.

**Acknowledgements.** I would like to thank many persons that helped me to reach this very important goal of my life in different ways. From the professional perspective, I thank my supervisor Prof. Vittorio Murino and the person I have been working with in these years, Dr. Marco Cristani. My thanks goes also to the colleagues and friends of the VIPS laboratory, where I spent a lot of hours. A special thank goes to the guys from the University of British Columbia, that guested me for six month and helped me a lot in my professional growth. In particular, Prof. Nando de Freitas was patient with me in explaining every theoretical issue in very basic detail. And also he introduced me to the Machine Learning community.

I also would like to thank who funded this research. First of all, I thank for the grant I received from the *Dipartimento di Informatica, Università di Verona*. Without that important funds, probably I would never join this PhD program. I thank the european project I was involved in, that is, EU-Project FP7 SAMURAI, grant FP7-SEC- 2007-01 No. 217899. It helped me to travel around the world, present the papers at the conferences and attend very interesting summer schools. The work I did at UBC was also partially supported by CIFAR's NCAP program and NSERC. I thank the organizers of ICML to provide a scholarship for traveling to the US, so I could attend the conference and present the work on attentional tracking. The work on person re-identification was partially supported by IBM, through the IBM Best Student Paper Award track: Computer Vision, I received at ICPR.

Most important was the personal support that the persons gave me during these not so simple years. I thank my fiancé Annarita to support me during this hard travel to reach the goal. She had a lot of patience especially during the nights I spent working during the last days of the conference deadlines. I also thank my parents Lucia and Alberto to help me economically and morally in all my life. My brother Johnny and his girlfriend helped me to not give up the studies. Without all of them, I think this important goal of my life would not have been possible.

**List of publications.** Here a complete list of publication I have been involved during my thesis:

[27] L. Bazzani, D. Bloisi, and V. Murino. A comparison of multi hypothesis kalman filter and particle filter for multi-target tracking. In Performance Evaluation of Tracking and Surveillance workshop at CVPR, pages 47–54, Miami, Florida, 2009.

[28] L. Bazzani, M. Cristani, M. Bicego, and V. Murino. Online subjective feature selection for occlusion management in tracking applications. In 16th IEEE International Conference on Image Processing (ICIP), pages 3617 –3620, November 2009.

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

[30] L. Bazzani, M. Cristani, G. Pagetti, D. Tosato, G. Menegaz, and V. Murino. Analyzing groups: a social signaling perspective. In Video Analytics for Business Intelligence, Studies in Computational Intelligence. Springer-Verlag, 2012.

[32] L. Bazzani, M. Cristani, A. Perina, and V. Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recognition Letters, 2011.

[31] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by hpe signature. In 20th International Conference on Pattern Recognition (ICPR), pages 1413–1416, August 2010. IBM Best Student Paper Award track: Computer Vision.

[33] L. Bazzani, N. de Freitas, H. Larochelle, V. Murino, and J-A Ting. Learning attentional policies for object tracking and recognition in video with deep networks. In International Conference on Machine Learning (ICML), 2011.

[34] L. Bazzani, N. de Freitas, and J-A Ting. Learning attentional mechanisms for simultaneous object tracking and recognition with deep networks. In Deep Learning and Unsupervised Feature Learning Workshop at NIPS, Vancouver, Canada, 2010.

[35] L. Bazzani, V. Murino, and M. Cristani. Decentralized particle filter for joint individual-group tracking. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2012.

[36] L. Bazzani, D. Tosato, M. Cristani, M. Farenzena, G. Pagetti, G. Menegaz, and V. Murino. Social interactions by visual focus of attention in a three-dimensional environment. Expert Systems, 2012.

[53] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In British Machine Vision Conference (BMVC), 2011.

[62] M. Cristani, L. Bazzani, G. Pagetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of f-formations. In British Machine Vision Conference (BMVC), 2011.

[63] M. Cristani, G. Pagetti, A. Vinciarelli, L. Bazzani, G. Menegaz, and V. Murino. Towards computational proxemics: Inferring social relations from in-

terpersonal distances. In International Conference on Social Computing (SocialCom), 2011.

[67] M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas. Learning where to attend with deep architectures for image tracking. Neural Computation, 2012.

[76] M. Farenzena, L. Bazzani, V. Murino, and M. Cristani. Towards a subject-centered analysis for automated video surveillance. In Proceedings of the 15th International Conference on Image Analysis and Processing (ICIAP), pages 481–489. Springer-Verlag, 2009.

[77] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2360–2367, June 2010.

[78] M. Farenzena, A. Tavano, L. Bazzani, D. Tosato, G. Pagetti, G. Menegaz, V. Murino, and M. Cristani. Social interaction by visual focus of attention in a three-dimensional environment. In Workshop on Pattern Recognition and Artificial Intelligence for Human Behavior Analysis at AI*IA, 2009.

Note that not all these publications will be presented in the thesis. Only the papers I have contributed more will be discussed.

**Publications per Topic.** In the following, the list of topics presented in this thesis with associated the publications that I was involved:

- Target tracking [27, 28],
- Attentional policies for tracking [33, 34, 67],
- Person re-identifications [31, 32, 53, 77],
- Social interaction discovery [30, 36, 62, 63, 76, 78],
- Tracking of social interactions [29, 35].

# Contents

**Part I Visual Tracking**

## Part II Person Re-identification for Multi-camera Tracking

## Part III Analyzing Groups

**Part IV Conclusions**

# 1

# Introduction

Every day millions and millions of Closed-Circuit TeleVision cameras (CCTV) monitor environments, collect and record data. The recorded data are per se not useful for the humans. They should be "manually" attended and analyzed in real-time by human experts all day long to extract high-level information of interest. Think that only United Kingdom has at least 4.2 million of cameras, it turns out that there is one camera for each 14 citizens[1]. Let us assume for now that a single human operator can monitor more cameras at the same time. However, this is not possible due to several limitations of our visual system. Neuroscience models agree with the theory that humans can only focus the attention in particular regions of the whole field of view at a certain instant. Thus, in a real situation where the human expert has many and many screens to attend, it is almost impossible to be aware of every single event that happens in a certain scenario. From this reasoning, we deduce that for a proper monitoring of an environment we would need a single human operator for a single camera (or at most few cameras). This means that one person over 14 has to be an human expert that monitors one camera. But, this is not a plausible option in a real scenario. This strongly highlights the need of *automatic* video surveillance systems that analyze the data and extract some high-level information to direct the attention of the human expert only when it is strictly necessary, in order to decrease the false negative and false positive rate.

The standard monitoring systems raise also ethical problems concerning the abuse of these technologies. The human expert could intentionally use the monitoring system to invade the privacy of the people. Or, worse this technology could be used improperly for not so genuine purposes. This is an additional motivation for having automatic video surveillance systems. The system will enable the expert to access to the data only when the detected situation is serious enough (*e.g.*, an abnormal event).

Several aspects concern to the fulfillment of an automatic video surveillance system. These aspects are mostly related to computer vision, machine learning and pattern recognition research topics. In this thesis, we will care more about these technical, research aspects rather than the ethical and philosophical point of views. In particular, there are several recent european project that treats these problems

[1] This analysis has been done in 2002 `http://en.wikipedia.org/wiki/Mass_surveillance`.

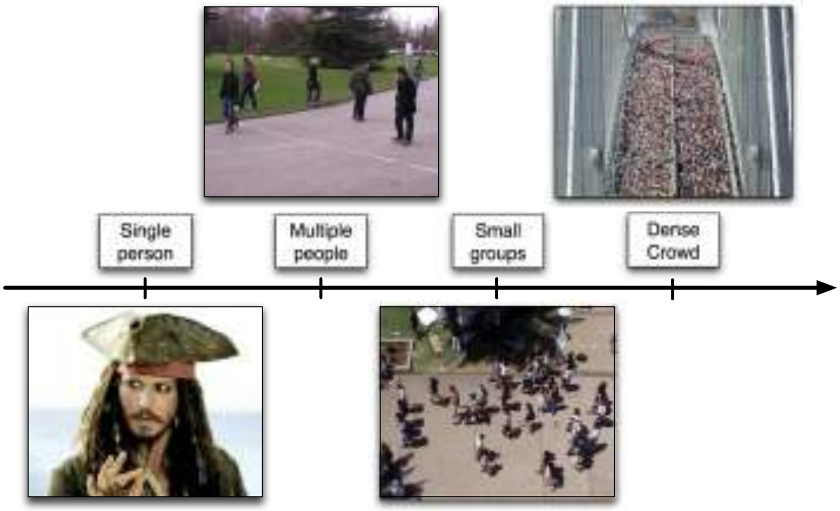


**Fig. 1.1.** Level of analysis when increasing the number of monitored people.

(for example, SAMURAI[2]). The general purpose of this thesis is to investigate models for automatic surveillance that extract some high-level information from the data.

In automatic surveillance, several levels of analysis can be recognized in two orthogonal dimensions as shown in Fig. 1.1. and Fig. 1.2. The first dimension (Fig. 1.1) takes into consideration the number of people that appear in the camera-monitored environment. The simplest level of video surveillance is to monitor a *single person*. A lot of work has been done in this direction. Face detection and tracking is usually the most popular application. The second level of analysis is characterized by the presence of *multiple people*. In this case, the problem becomes harder, because several person can disappear for a long period and appear again after an occlusion for example. In the third level, we find a scenario where people naturally form *small groups*. Since humans are "social animals", they usually travel, walk, go to the cinema, work, *etc.* with other people. An unpublished study by McPhail found that 89% of people attending an event came with at least one other person [92, 93]. Thus, small groups are very frequent in surveillance scenarios. The last level is called *crowd analysis*, where the surveillance scenario is very crowded. Imagine scenes like the New York City marathon or pilgrims circling around Kabba in Mecca. These are typical scenarios where a single-person analysis is hard to perform.

In this thesis, we analyze several aspects in this dimension. We analyze the single-target scenario, the multi-target scenario and the small group scenario. Crowd analysis is also very interesting, but it is out of the scope of this thesis,

[2] SAMURAI: Suspicious and Abnormal behavior Monitoring Using a netwoRk of cAmeras for sItuation awareness enhancement http://www.samurai-eu.org.

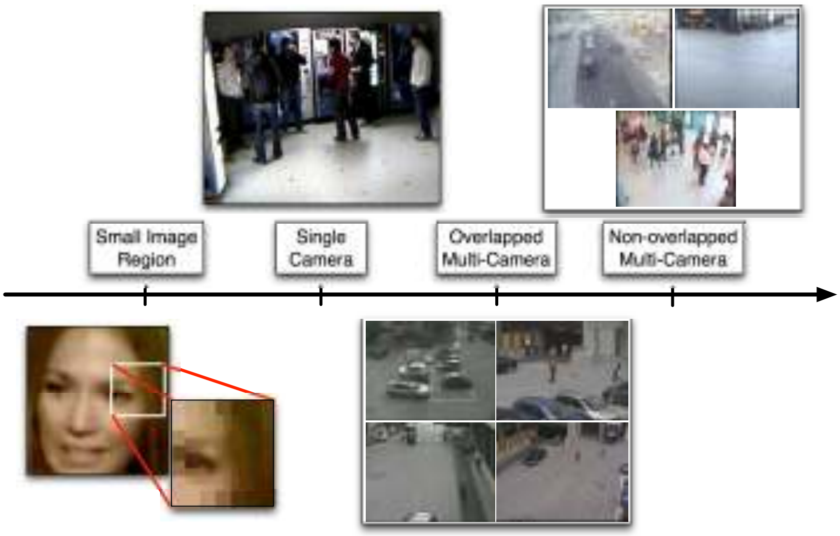


**Fig. 1.2.** Level of analysis when increasing the number of observation that the system gathers.

because it could be itself a topic for a PhD thesis. We refer the readers interested in this topic to start from [9, 109, 211].

The orthogonal dimension (Fig. 1.2) concerns the number of observations that the system can gather at each time. Only *local information* can be available, that is, local patches in the image. In this case, we avoid to process the full image, but instead the model has a learning mechanism that automatically decides where is worth to look at. The second level concerns the *single-camera* analysis, where the algorithm input is a single or a sequential set of images recorded from a static (or moving) camera. Then, the natural extension is to the multi-camera setting. We distinguish two cases: *overlapped and non-overlapped cameras*. The former is usually simpler, because some geometrical cues can be used to estimate jointly the position of the people. However, in a real scenario (for example, a bank) it is common to have few cameras for each room with very little overlapping view and thus this kind of analysis is not very useful. On the other hand, the non-overlapped camera analysis is a bit harder, because targets travels between different cameras going across non-monitored areas, but a more realistic situation.

In this thesis, we propose models and methods to deal with local information, single camera and multiple non-overlapped camera. We focus on non-overlapped camera because it is usually a more general scenario. Moreover, the proposed methods can also be integrated with geometric constrains to deal with overlapped cameras.

Following the description of the Marr's seminal work [164], we analyze the problem from the perspective of abstraction levels. In our context, an abstraction level is defined depending on the meaningfulness of the description given by the

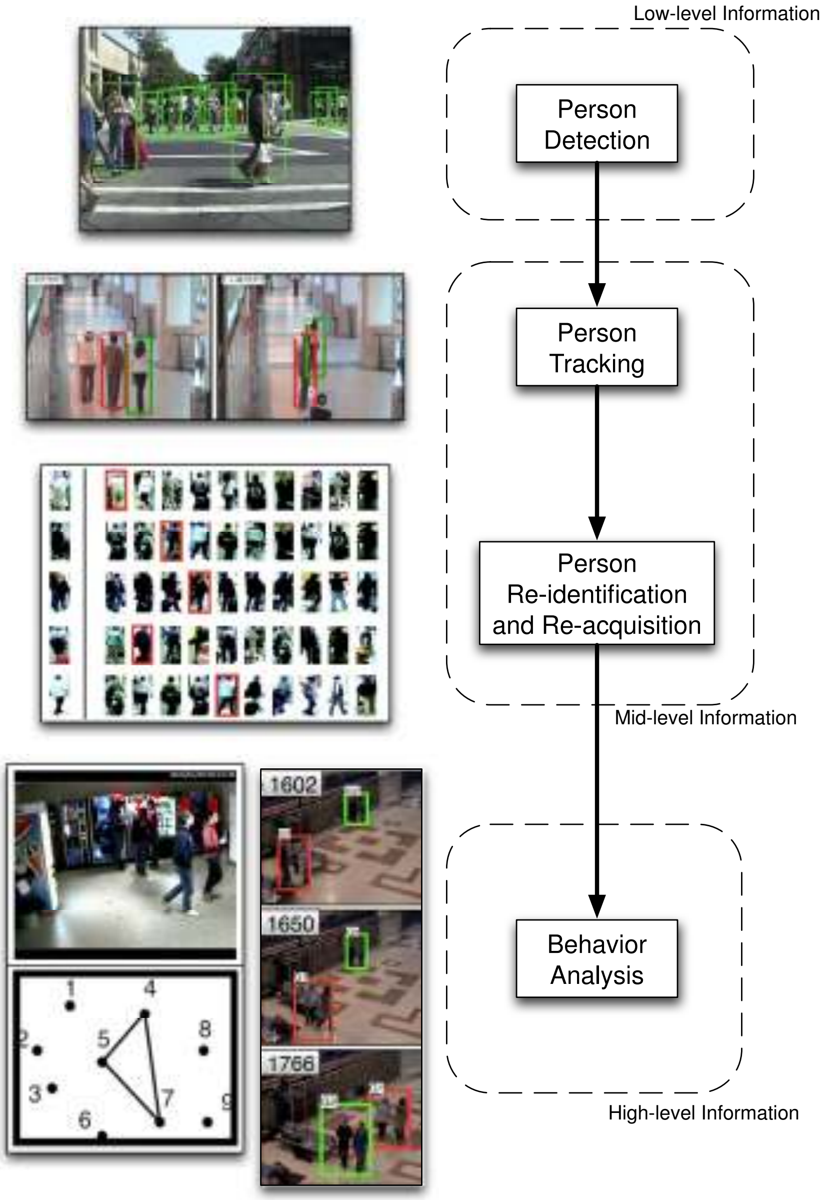


**Fig. 1.3.** Automatic surveillance can be broken down in subproblems: First, person detection and sometimes background subtraction are performed. The results are fed into a person tracker that connect the detections over time. Then a person re-identifier connect tracks over multiple cameras. Finally, the behavior analysis extracts some relevant high-level information given the results of the lower levels.

output of the algorithm. Likewise Marr, we define three levels: the low level gives a very raw description of the scene, the medium level extracts some information of interest that can be used by a human. Finally, the highest level provides very useful information about the scene and the actions that are happening, that can be easily used from every people. An instance of abstraction levels for the surveillance problem is depicted in Fig. 1.3. Depending on what level of description is required, a surveillance system can be broken down is the following sub-problems [117]:

- *Person detection.* The goal is to localize the persons of interest in the image usually with a bounding box, considering each frame independently (no temporal reasoning).
- *Person tracking* considers the temporal component to localize the persons of interest, taking as input the person detections and the image observations. Intuitively, when the person detector fails the tracking algorithm fills the empty frames by assuming a certain smooth dynamics of the person.
- *Person re-identification* consists in recognizing an individual in diverse locations over different non-overlapping camera views, considering a large set of candidates. An instance of this problem is *person re-acquisition*, that corresponds to re-identification in the same camera, for example, to overcome occlusion failures of the tracker.
- *Behavior analysis* detects some behaviors of interest from the camera-monitored scenario. The definition of behavior is ambiguous and it could have several interpretations, that depends on the application. For example, in the video surveillance context behavior analysis could mean: detect when an abnormal event occurs; or recognize the different human actions or activities; or detect social interactions between subjects. Due to the generality of its meaning, a lot of techniques and applications have been studied in the last decade.

In this taxonomy, this thesis is transversal. We analyze person tracking, re-identification and re-acquisition, and some interesting aspects of behavior analysis concerning small groups of people. A further level of taxonomy can be defined by analyzing the level of information that the sub-problem output provides to the human operator: person detection provides *low-level information*, because it is hard to make consideration from only results on still images. Person tracking and re-identification connect temporal data and different cameras, respectively. For this reason, they extract *mid-level information.* Behavior analysis brings out *high-level information*, because it provides some high-level description of the dynamics in the monitored environment. The only information that usually helps the human expert during a surveillance task is the last, but it is worth noting that it strictly depends on the analysis at the lower levels. In these last years, the trend goes towards the high-level analysis, however the lower levels still need some attention from the research point of view.

Given this brief introduction of the context that this thesis is going to investigate, it is important now to understand which mathematical tools fit with the problems we deal with and how this thesis extends them. In section 1.1 we give a brief mathematical introduction of the core model of the thesis, and we introduce the novelties proposed in this thesis (please the reader can refer to the respective chapters for the details). In section 1.2 we summarize the main contributions of the work. Section 1.3 concludes the chapter giving an overview of the organization of the thesis.

## 1.1 State-space Model and its Extensions

This work intends to go beyond the standard multi-target tracking, hence the formalization of the basic problem is necessary to understand the extensions we propose in this thesis. Therefore, we present the state-space models used for target tracking. This modeling will follow the reader across all the thesis, and for this reason it is an essential material that has to be deeply understood. We first analyze the standard state-space model for tracking, and then we will describe how this thesis intends to extend the model from different perspectives and for the different purposes mentioned in the previous section.

Target tracking is an example of sequential data analysis that in general can be represented in terms of state-space models [172]. The state-space models have several advantages with respect to the classical heuristic approaches: 1) the prediction of the future does not depend on a limited temporal window, 2) incorporating prior knowledge of the problem is easier and 3) they do not have problems with multi-dimensional data.

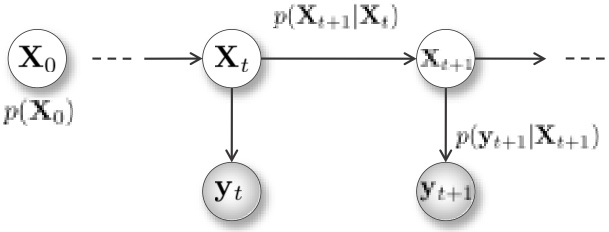


**Fig. 1.4.** Standard Tracking Model. The building block for the models proposed in this thesis.

The state-of-the-art state-space model for target tracking is shown in Fig. 1.4. Actually, Fig. 1.4 is very general and represents a family of models (depending on the assumptions). In our work, we focus on Markovian, nonlinear, non-Gaussian models. In this setting, the unobserved/hidden signal (object's position, velocity, scale, orientation or discrete set of operations) is denoted as $\mathbf{X}_t \in \mathbb{R}^{n_x}$, where $t \in \mathbb{N}$ represents the discrete time and $n_x$ is the size of the state space. This signal has an initial distribution $p\left(\mathbf{X}_0\right)$ and a dynamical model $p\left(\mathbf{X}_{t+1}|\,\mathbf{X}_t\right)$. The signal has an observable part, denoted as $\mathbf{y}_t \in \mathbb{R}^{n_y}$, that is also called the current measurement. The relation between the state $\mathbf{X}_{t+1}$ and the observation $\mathbf{y}_{t+1}$ is modeled by the observation distribution $p\left(\mathbf{y}_{t+1}|\mathbf{X}_{t+1}\right)$. We denote as $\mathbf{X}_{0:t}$ and $\mathbf{y}_{1:t}$ the sequence of states and measurements up to time $t$, respectively. The split between hidden and observable variables enables us to define the problem in terms of graphical models (Fig. 1.4).

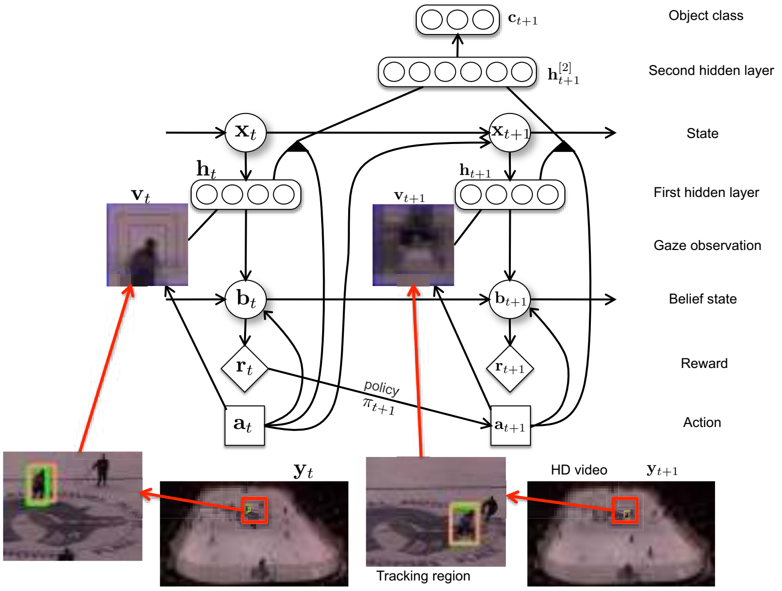


**Fig. 1.5.** Attentional Model for Tracking and Recognition..

The main goal in image tracking is to infer (estimate) the filtering distribution $p(\mathbf{X}_t|\mathbf{y}_{1:t})$ (over most recent state) or the posterior distribution $p(\mathbf{X}_{0:t}|\mathbf{y}_{1:t})$ (over a whole trajectory). Eventually, the information that is worth to extract is the state estimate, defined as the expected value of $\mathbf{X}_t$ under the posterior distribution, that is, $\tilde{\mathbf{X}}_t = \mathbb{E}_{p(\mathbf{X}_t|\mathbf{y}_{1:t})}[\mathbf{X}_t]$. In the next chapter, we will see more details about this model, for example, how to perform approximate inference in the probabilistic model of Fig. 1.4 using sequential Monte Carlo methods (like particle filtering).

Let us analyze the extensions of the state-space model we proposed in this thesis. The first work we present is focused in the definition of a robust observation distribution $p\left(\mathbf{y}_{t+1}|\mathbf{X}_{t+1}\right)$ such that similar targets are tracked even if they are close, partially occluded and hardly discernible in appearance. To this end, in Chapter 2 we proposed an online subjective feature selection mechanism embedded into the model of Fig. 1.4.

In Chapter 3, the model is extended to deal with attentional policies for tracking and recognition. In this work, only a very small portion of the image is available at each time (gaze data). The goal is 1) to select the region that is best to attend to, and 2) to gather information over time in order to perform robust recognition. The model is an extension of Fig. 1.4 and it is presented in Fig. 1.5. It mainly differs from the standard model for tracking, because has a online learned control mechanism that selects regions in the image, it relies on deep models [107, 203] and it performs also recognition. In addition, it has affinities with the cognitive models in neuroscience [207], because point 1) and 2) can be seen as main characteristics our attentional system. We will make more explicit this peculiarity in Chapter 3.

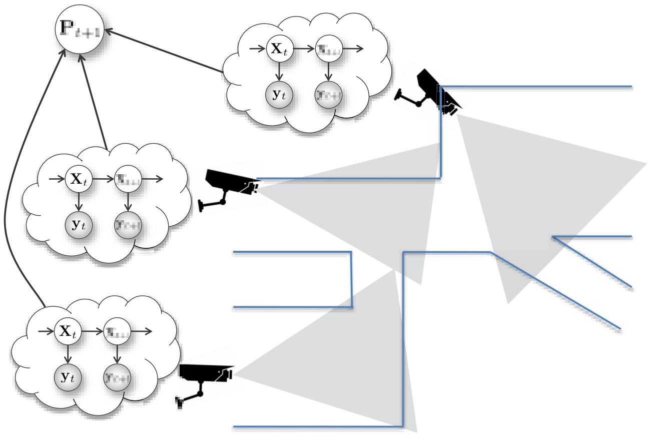
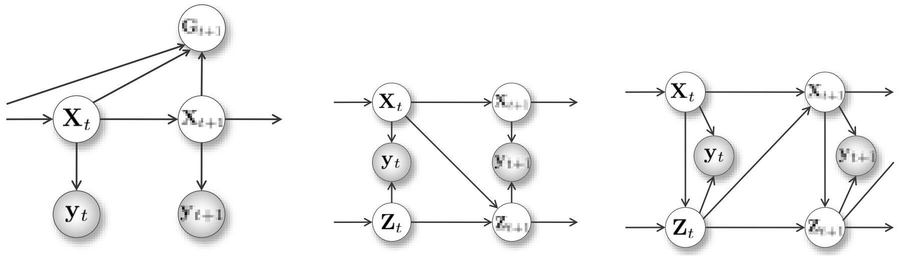


**Fig. 1.6.** Multi-camera model for person re-identification and re-acquisition. The block of Figure 1.4 is extended to multiple cameras. Every camera has its own tracker, the connection between state estimates is performed at higher level $\mathbf{P}_{t+1}$.

It is also worth investigating what happens when dealing with a camera network. In particular, when multiple independent tracking models are available (Fig. 1.6), we need also to link the results across different cameras in order to keep persistent identifiers in the camera network (the variable $\mathbf{P}_{t+1}$ in Fig. 1.6). In the literature, this problem is called person re-identification. In Chapter 4, we propose a standard pipeline for the problem. The methods we presented takes advantage from the sequential nature of the data gathered by trackers and from the natural symmetries that characterizes the objects, in order to make more robust the matching between individuals descriptors without using any geometrical constraint/reasoning. Moreover, the same methods can be used for person re-acquisition when a track is lost due to strong occlusions.

The behavioral aspect is the core part of the thesis. It is extensively analyzed from two points of view: 1) detection of social interactions in video focusing on small groups and 2) (joint) tracking of groups and individuals. The former (Fig. 1.7(a)) takes the tracker results as input to infer some high-level information ($\mathbf{G}_{t+1}$ in Fig. 1.7(a)) by gathering information over time. The latter (Fig. 1.7(b-c)) extends the probabilistic framework of standard tracking to deal with both the individual and the group tracking issue. The proposed models will be discussed in Chapter 5 and Chapter 6, respectively[3].

[3] Note that the models of Fig. 1.6 and 1.7(a) are not formal graphical models. In fact, inference of tracking and re-identification or groups is not performed jointly. First, we estimate the tracks and then we estimate the second level (re-identification or groups)



(a) Social Interaction (b) Co-PF (c) DEEPER-JIGT

**Fig. 1.7.** Model of (a) social interaction detection and (b-c) group tracking. The block of Figure 1.4 is extended (a) to gather information other time in order to discover groups in the scene, (b-c) to deal with the group tracking (estimate of $\mathbf{Z}_{t+1}$) in independent and joint way, respectively.

## 1.2 Contributions

Summarizing, the main contributions of this thesis divided by chapters are:

- **Tracking with online subjective feature selection, [Chapter 2],** to handle partial occlusions especially when the targets are hardly discernible in appearance. We propose a novel observation model for the particle filtering framework that highlights and employs the most discriminant features characterizing a target with respect to the neighboring targets.

- **Attentional models for tracking and recognition, [Chapter 3].** Motivated by cognitive theories, the model consists of two interacting pathways: ventral and dorsal. The ventral pathway models object appearance and classification using deep networks. At each point in time, the observations consist of retinal images, with decaying resolution toward the periphery of the gaze. The dorsal pathway models the location, orientation, scale and speed of the attended object. The posterior distribution of these states is estimated with particle filtering. Deeper in the dorsal pathway, we encounter an attentional mechanism that learns to control gaze so as to minimize tracking uncertainty.

- **Person re-identification and re-acquisition, [Chapter 4].** We present a pipeline for person re-identification and three appearance-based descriptors for recognizing an individual in diverse locations over different non-overlapping camera views or also the same camera, considering a large set of candidates. They consist in the extraction and matching of features that model complementary aspects of the human appearance. Features from multiple images of the same individual are fused to achieve robustness against very low resolution, occlusions and pose, viewpoint and illumination changes. We also show that if the images are segmented in body parts exploiting symmetry and asymmetry perceptual principles, the descriptor gains even more robustness. Moreover, we prove that the proposed descriptors can be used for modeling the target appearance in tracking applications.

- **Model for social interaction detection from video, [Chapter 5].** We introduce basic methods to deal with groups of people in surveillance settings, that embed notions of social psychology into computer vision algorithms, thus offering a new research perspective in the social signal processing. Methods to detect groups of subjects and to infer how people and groups interact are proposed. In particular, the inter-relation pattern matrix and the subjective view frustum are introduced for the first time here (and the related published work).
- **Tracking of social interactions, [Chapter 6].** Two complementary solutions are proposed to deal with this problem: collaborative individual-group tracking and joint individual-group tracking. The former assumes that individual and group tracking are two independent problems that probabilistically share joint information in a collaborative way introducing some approximation. Instead, the latter deals with the problem in a joint way. The full joint posterior probability distribution is estimated. Therefore, the collaborative framework becomes more probabilistic elegant.

## 1.3 Organization of the Thesis

The thesis is organized in three main, logical parts. The first part that includes Chapter 2 and 3 analyzes the tracking problem from two different points of views: the engineering one (Chapter 2) and the perceptual one (Chapter 3). The former is a class of solutions usually aimed at solving specific problems (*e.g.*, object tracking) to reach the state-of-the-art results. The latter instead takes inspiration from how the human perceptual system works, trying also to understand how the brain works. In Chapter 2, the standard probabilistic framework of dynamic state estimation for tracking is first presented. We also present an online subjective feature selection for tracking to handle partial occlusions. The chapter presents the papers published in [27, 28]. The perceptual point of view of tracking is discussed in Chapter 3. Taking inspiration by theories of the visual system, we propose a novel model for tracking and recognition. The papers described in this chapter are [33, 34, 67].

The second part of the thesis that includes Chapter 4 focuses on the extension of tracking when dealing with multi-camera networks. In particular, we cope with the problem by proposing 1) a full person re-identification pipeline and 2) three appearance-based descriptors for human modeling, in contrast with the standard geometric reasoning that helps only in the overlapped cameras scenario. We present three descriptors and the matching procedure for person re-identification and re-acquisition discussed in the papers we published in [31, 32, 77].

The last part that includes Chapter 5 and 6 that cover some aspects concerning social interaction discovery with particular focus on groups of people. First, Chapter 5 describes how to detect interactions between individuals in groups. The work has been published in [30,36,76,78]. Second, Chapter 6 focuses on individual-group tracking. In this chapter, we will present the two different solution published in [29] and [35].

Finally, Chapter 7 draws the conclusions of the thesis. We will discuss the challenges that each the proposed models and applications presents and what are the future directions for improvements. Note that each chapter contains the state of the art of the problem at the beginning and a conclusion in the end. We preferred to keep them spread on the thesis instead to concentrate everything in a single chapter, in order to focalize better the attention of the reader on a single topic at time by presenting incrementally the models.

# Part I

# Visual Tracking

# 2

# Multi-target Tracking

In the recent years, the interest of many researchers has been captured by the deployment of automated systems to analyze a continuous streaming of data, such as, videos. This is a consequence of the fact that video cameras are installed daily all around the world, for surveillance and monitoring purposes. In fact, it is essential to built automatic algorithms that process the large amount of data, minimizing the interaction with the human. Many applications can profit from this automatic analysis, such as video surveillance, sport analysis, smart rooms, social interaction analysis, quality control analysis, and many others. In this thesis, we mainly focus on the video surveillance application, but note that part of the techniques discussed here can also be used for other purposes.

As we have already discussed in Chapter 1, a multi-camera system that has to deal with sequential data has to face several problems: 1) person detection, 2) person tracking, 3) person re-identification and 4) behavior analysis. The aim of this section is to describe and discuss the problem of multi-person tracking. In general, object tracking aims at localizing one or more objects of interest in the space-time domain of a video sequence. In practice, we want an algorithm that localizes the objects in each image of the sequence and that connect them through different time steps. This cannot be done using just an object detector (for details see Sec. 2.1), because we need to consider the non-linearity and noisy nature of the problem. Thus, we need a tracking method that copes with this issues.

Even if tracking is one of the most investigated problems in the last decade from both the signal processing and the computer vision communities, it is still far from being definitively solved. A lot of challenging issues are involved in object tracking, such as multiple targets, objects hardly discernible, objects similar to the background, shadows, crowded scenarios, occlusions, illumination changes, viewpoint variations, non-rigid objects, appearance changes, non-linear dynamics, appearance model drifting, *etc.* The *universal* tracker has to be robust to all these issues to avoid failures. However, most of the work done so far tackles only a subset of them, because it is very hard to deal with all of them jointly. In the same vein, the method proposed in this chapter deals with few sub-problems. In particular, the main objective is to propose a method that deals with partial occlusions between hardly discernible targets, because usually occlusion is one of the hardest problems in tracking. Thus, we introduce a mechanism for *online subjective feature* *selection*, that selects and exploits the most discriminant features characterizing a single target with respect to the neighboring objects in order to better deal with partial occlusions when the involved objects are similar (Sec. 2.3).

In this chapter, we will first present the state of the art of object (person) tracking (Sec. 2.1). We give a general background of the problem because it is important to know the literature, in order to fully understand also the following chapters that describe the work *beyond target tracking*. In Sec. 2.2, we present the particle filtering approach because it is one the most interesting, effective and efficient approaches for tracking. In Sec. 2.3, the main contribution of this work is described. Then, the experimental trials are presented in Sec. 2.4 to test the techniques discussed. A final discussion concludes the chapter in Sec. 2.5.

## 2.1 Related Work

An overview of the state of the art of the tracking techniques are presented in this section. We will focus on some of the most interesting aspects of the problem: the different tracking approaches (filtering and data association, and tracking-by-detection techniques), the appearance representation of the target and its updating, and the occlusion issue, that are, the main background that we need in this and next chapters.

**Filtering and Data Association.** Several approaches have been proposed for object tracking. Probably the most interesting ones have been borrowed from the signal processing community [21,22,70,124], where tracking is formalized as a discrete non-linear system with non-Gaussian noise, *i.e.*,

$$\begin{aligned}\mathbf{X}_{t+1} &= f(\mathbf{X}_{0:t}, \xi^{\mathbf{x}}_{t+1})\\ \mathbf{y}_{t+1} &= h(\mathbf{X}_{t+1}, \xi^{\mathbf{y}}_{t+1})\end{aligned}$$

where $\mathbf{X}_{t+1}$ is the unknown state of the system (*e.g.*, the position) at time $t+1$, $\mathbf{X}_{0:t}$ are the estimated states up to the current time, $\mathbf{y}_{t+1}$ is the observable measurement at time $t+1$, $\xi^{\mathbf{x}}_{t+1}$ and $\xi^{\mathbf{y}}_{t+1}$ are noises associated to the dynamics and the measurement, and $f(\cdot)$ and $h(\cdot)$ are unknown non-linear function that governs the dynamics of the state and the measurement process, respectively. The goal is to estimate the state of the system $\mathbf{X}_{t+1}$ given all the measurements up to the current time $\mathbf{y}_{1:t+1}$. In other words, we want to filter out the correct state from the noisy observations. For this reason, this problem takes the name of *filtering* problem.

Several assumptions have been made on the above formulation in order to make the problem tractable. Kalman Filter models [124] represents the optimal solution with the assumptions of linear equations and Gaussian noise. Relaxing the linearity assumption, a sub-optimal solution can be found in the the Extended Kalman Filter [21] and the Unscented Kalman Filter [120,148,157]. However, it is possible to prove that these methods can deal with problems with a degree of non-linearity not too high, because of the approximations. A general solution (*i.e.*, without too strong assumptions) is given by particle filter that deals with the non-linearity and non-Gaussianity of the system [12, 70]. Among the realm of the tracking strategies, an important role is played by the particle filtering [70] for several reasons: 1) its general framework, *i.e.*, it deals with non-linear, non-Gaussian state spaces, 2) and its simplicity, *i.e.*, you can implement it in 10 minutes with few lines of code, and 3) its efficiency, *i.e.*, it can be implemented for real-time applications.

Particle filtering offers a probabilistic framework for recursive dynamic state estimation. It is a recursive inference procedure composed by three steps: 1) in the *sampling* step, hypotheses which describe the state of the system are generated from a candidate probability distribution (proposal distribution); the posterior distribution over the states is approximated by a set of weighted samples, where each sample is an hypothesis whose weight mirrors the associated probability. 2) In the *dynamical* step, each hypothesis is moved accordingly to a certain dynamical model. And 3) in the *observational* step the hypotheses that agree at best with an observation process are awarded so avoiding a brute-force search in the prohibitively large state space of the possible events.

Particle filtering was born in computer vision for single-target tracking with CONDENSATION [112], and later was extended to a multi-target tracking scenario with BraMBLe [113]. Multi-target particle filters follow different strategies to achieve good tracking performances avoiding huge computational burdens. These are due primarily to the high number of particles required to explore the state space, which is (in general) exponential in the number of targets to track. In particular, several techniques to deal with multiple targets have been discussed in literature: an independent particle filter for each target with sequential importance sampling [70] allows to sample in independent state spaces [43, 44, 50] (one space for each target); a single filter defined in the joint state space [113, 128] where sequential importance sampling or Markov Chain Monte Carlo (MCMC) can be used [70]. Recently, an hybrid, interesting solution has been introduced in [142], that is called the Hybrid Joint-Separable filter. It maintains a linear relationship between the number of targets and the number of particles by sampling in the independent state spaces and considers dynamics and observation in the joint space. This enables us to sample in the independent spaces and also to model occlusions and interactions between targets in the joint space (see Sec. 2.2.1 for details).

The general tracking problem concerns with multiple measurements, therefore each target needs to be validated and associated to a single measurement in a *data association* process [21]. Nearest Neighbor [21, 151] and Probabilistic Data Association (PDA) methods [94, 205] deal with single targets-multiple measurements data association problem, while Joint Probabilistic Data Association (JPDA) [22] and multiple hypothesis methods [41] deal with multiple targets-multiple measurements data association problem. Those methods are usually combined with Kalman filter, *e.g.*, nearest neighbor filter [151], multi-hypothesis tracker [41], and JPDA filter [22]. The drawback of those strategies is that they relies on the assumptions of linearity and Gaussianity of the Kalman filter and this it cannot manage complex scenarios. Techniques that combines data association methods with particle filtering to accommodate general non-linear and non-Gaussian models are MCMC DA [129, 274], Monte Carlo JPDA [248], Independent Partition particle filter [248], and Joint Likelihood filter [205]. Other advanced techniques can be used for data association. One example is [191], where data association is improved by the knowledge groups and the joint modeling of pedestrian trajectories using a conditional random field.

Plenty of techniques have been proposed in literature, however, in visual tracking applications it is used to choose the simplest methods that lead to a low computational burden, that is, particle filtering combined with nearest neighbor data association or PDA. Successful recent tracking systems like [43,44] exploits particle filter with nearest neighbor DA. In practice, a greedy approach for DA is often sufficient, as pointed out by [265]. Other interesting combinations are possible, such as: the Unscented particle filter (particle filtering that uses UKF as proposal distribution) [245], the particle Probabilistic Hypothesis Density (PHD) filter (particle filter applied to the random sets theory) [57, 160], the (Reversible-Jump) MCMC particle filter [23, 128, 278] (deals better with high-dimensional spaces) and so on. We are not going to describe them in details because it is out of the scope of this thesis, but the reader can start from the cited papers if interested.

**Tracking-by-detection Methods.** In visual tracking, depending on the measurements $\mathbf{y}_t$ that one decides to use, we can define *detection-free* and *tracking-by-detection* methods. The former class of approaches employs directly the entire image or its feature representation as measurements (we discuss it in the "Appearance Representation" subsection), instead the latter class uses detections given by an object detector [68,80,82,126,253]. Tracking-by-detection methods uses detections directly as input, but hybrid approaches are usually preferred, *i.e.*, using detections and features extracted from the image[1]. The strategies to combine trackers and detectors are manifold: 1) perform data association with detections [43,44], 2) embed them into the observation model in particle filtering [43,44], 3) use them to generate new hypotheses in particle filtering [50, 178] and 4) associate detections over time to perform *deterministic* tracking [110, 135, 152].

Detections can be used for data association [43, 44]. Given a set of detections and a set of tracking estimates, the aim is to find the detections associated to each track. In [44], a greedy algorithm is proposed to compute a similarity matrix between detections and tracks defined in terms of appearance similarity, a gating function and a classifier score. The correspondence problem is solved using the Hungarian algorithm on the similarity matrix. The same authors have also proposed to use the detection confidence score given by a detector in the observation model of the particle filter. In this way, the algorithm avoids to consider background hypotheses that are hardly discernible from the target and that could drift the track away.

Instead, the approach proposed in [50, 178] modifies the proposal step of the particle filtering approaches. The proposed method aims to generate hypotheses with high probability where detections are available. They define the proposal distribution as a combination of the standard dynamical proposal and a detection-based proposal. The combination between those proposals enables the algorithm to be robust also in case of false negatives and positives. This approach is described more in details in Sec. 2.2.2 because in practice it turns out to be very effective.

[1] We still refer to these hybrid methods as tracking-by-detection.

Finally, associating detections over time to perform deterministic tracking has been proposed in [110, 135, 152]. Similarly, to the detections-for-data-association framework (point 1 of the list) where detections were associated to tracks, their method hierarchically associate detections at the current time with other detections at the previous time steps creating *tracklets*. A tracklet is a set of detections that represents the same individuals. Note that tracklets are different from tracks because a set of tracklets of the same object is a fragmentation of the track with some missing parts. For this reason, they propose a supervised learning algorithm to connect tracklets. Unfortunately, the method has two drawbacks: 1) it requires to collect and annotate pairs of tracklet examples in advance to train the classifier and 2) it is not genuinely on-line because the algorithm has to accumulate tracklets in a time window.

**Appearance Representation.** We focus the discussion now on the features for object representation commonly exploited in tracking. Roughly speaking, such representations should be robust to hard recording conditions (*e.g.*, low resolution and scarce illumination). In addition, they have to be computationally efficient in order to comply to the large number of hypotheses that a particle filter has to evaluates at each time step. We follow the scheme proposed by [271], discussing first the data structures useful to represent objects, and then specifying the most common features employed.

Points are the poorest object representation, which are suitable for modeling targets that occupy small regions in an image, with little overlap. The object can be a single point (the centroid) [247], or a set of sparse points [225]. Covariance matrices of elementary features have been recently adopted to deal with non-rigid objects under different scales [196]. Geometric shapes as rectangle or regular ellipses serve to primarily model simple rigid objects affected by translation, affine, or projective (homography) transformations [61]. Elementary shapes may be employed to encode different body parts, such as head, torso and legs [113,142]. Patches may also be employed, to track salient parts of a target [136]. The contour representation, defining the boundary of an object containing its silhouette, can be suitable for tracking complex nonrigid shapes [272]. Articulated shape models as pictorial structures are composed by body parts held together with joints [10, 11, 81]. Such structures, used for human body pose estimation, essentially rely on two components, one capturing the local appearance of body parts, and the other representing an articulated body structure. Inference in a pictorial structure involves finding the *maximum-a-posteriori* spatial configuration of the parts, *i.e.* the body pose. Skeletal models [48,244] can also be extracted by considering the object silhouette, and can be used to model both articulated and rigid objects.

There are many appearance features for objects and the most employed are represented under the form of probability densities. They can be either parametric, such as Gaussian distributions [277] or mixtures of Gaussians [187], or nonparametric, such as Parzen windows [73] and histograms [61, 194]. The probability densities of object appearance features (color, texture) can be computed from the image regions specified by the shape models, *i.e.*, the internal region of an ellipse, a box, or a contour. Templates are formed using simple geometric shapes or silhouettes that model the whole targets or a portion of them [83, 113, 142, 178]. Their advantage is primarily due to the fact that they carry both spatial and appearance information, however, they can only encode the object appearance generated from a single view. Thus, they are only suitable for tracking objects whose poses do not vary considerably during the course of tracking. Active appearance models are generated by simultaneously modeling the object shape and appearance [71]. In general, the object shape is defined by a set of landmarks, and, similar to the contour-based representation, the landmarks can reside on the object boundaries or, alternatively, inside the object region. For each landmark, an appearance vector is stored which is in the form of color, texture, or gradient magnitude. Active appearance models require a training phase where both the shape and its associated appearance is learned from a set of samples using, for instance, principal component analysis [153]. The multi-view appearance models encode different views of an object. One approach to represent the different object views is to generate a subspace from the given views. Actually, subspace approaches such as principal component analysis or independent component analysis, have been used for both shape and appearance representations [40, 171].

Note that a weak appearance modeling of the target is not the only cause of tracking failure. Tracking may also fail when the object model is not properly updated, *i.e.*, the template update problem discussed in [166], and if a target becomes (even partially) occluded. We will briefly discuss these two problems in the following subsections.

**Template Update.** Many techniques that deal with the template update problem have been proposed in literature, known also as incremental learning or active appearance modeling. Here, we give a brief description of the main methods for the problem we are interested in; for a good overview the reader can refer to [166, 167].

One class of techniques are strictly based on the object representation. For example, the work presented in [153, 213] exploits subspace representations (principal component analysis) as appearance models. The main advantage of their technique is that does not require a training phase as in [40] but learns the eigenbases on-line during the object tracking process. Other works are based on the covariance object descriptor using the mathematical properties of the Riemannian manifolds [150, 183, 196, 258]. The update method proposed in [196] keeps a set of T previous covariance matrices and compute a sample mean covariance matrix on the manifold that blends all the previous matrices exploiting the Lie algebra characteristics. A similar strategy is followed by [258], but instead of computing the mean using Lie algebra they define they use the Log-Euclidean mean. In [183], an incremental technique is proposed: only the stored mean covariance matrix and the new covariance matrix are averaged.

The first class of approaches is bounded to the chosen object descriptor, however a more general solution is usually preferred. Machine learning techniques are general enough for being used for supervised [17, 18, 216, 217] and semi-supervised [100, 158] incremental learning of the template. In particular, the main goal of researchers of this field have been to exploit machine learning techniques and adapt them for the online learning for the tracking problem. This process gave birth to online multiple instance learning [17, 18], online random forest [216, 217], and online boosting [100, 158], among many others.

**Occlusion Handling.** The problem of the occlusions born from the nature of the pin-hole camera model, where the projective transformation maps the three-dimensional world to the two-dimensional image. It is clear that there is a loss of information due to the projection to a low-dimensional space. Occlusion is one of the major and challenging problem in applications of the computer vision, and also in multi-target tracking. In multi-target tracking, it corresponds to a loss of measurements for a certain period of time. Thus, that the algorithm should perform tracking without any clear observation of the target.

*Optical flow* can be used or estimated to detect occlusions [15] and depth ordering [237]. Joint estimate of occlusions and optical flow is posed as a convex minimization problem in [15]. In [237], the authors present a probabilistic model of optical flow in layers using temporal reasoning to capture occlusions, depth ordering of the layers and temporal consistency of the layer segmentation. However, these methods can only detect small occluded regions, because the assumption of optical flow is that temporal sampling has to be dense.

In tracking, *heuristics* are usually used to detect occlusions. The authors of [214] proposed a basic solution that compares the bounding boxes sizes and the last maximum sample likelihoods with recent historical values to detect when a target is occluded. However, this class of method usually can only work in case of short-term occlusions.

Another strategy to deal with occlusions is to not worry if a target is lost by the tracker, but try to reconstruct its trajectory when he/she is visible again. The goal is thus to associate tracklets trough time. We can also call this problem, *person re-acquisition*, and it can be dealt with using the person re-identification methods we propose in Chapter 4. Several methods that follow this idea have been proposed in literature (among them, [135, 152, 269]). However, these methods rely on the assumption that a long temporal window should be available. The second problem is that the method is not online: The occlusion is solved only when the association is performed. The authors of [273] avoid these problems by performing tracking-by-detection. They use a global optimization method (adaptive simulated annealing) to recover the track after inevitable tracking failures.

Other methods build an *occlusion map* exploiting or estimating some 3D information of the scene. For example, the camera is calibrated (extrinsic and intrinsic parameters) in [142]. Alternatively, one can estimate an homography between some planes in the scene and in the image to compute the depth ordering. Some geometrical characteristics of the scene can be used, for example, knowing the position of static objects in the environment [189]. If stereo cameras are available, the occlusion map can be easily estimated [75].

## 2.2 Tracking with Particle Filters

The standard approach to image tracking is based on the formulation of Markovian, nonlinear, non-Gaussian state-space models, which are solved with approx-

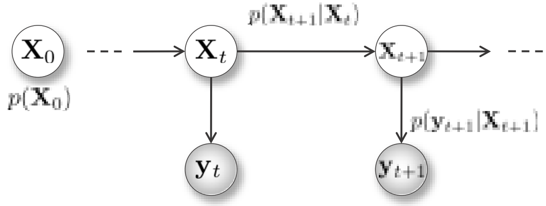


**Fig. 2.1.** Graphical model that represents the tracking problem. $\mathbf{X}_t$ is the unobserved signal (object's position, velocity, scale, orientation or discrete set of operations) at time $t$. $\mathbf{y}_t$ is the observed signal (image, detections, features, and so on) at time $t$. $p\left(\mathbf{X}_0\right)$ is the initial distribution, $p\left(\mathbf{X}_t|\,\mathbf{X}_{t-1}\right)$ is the transition model and $p(\mathbf{y}_t|\mathbf{X}_t)$ is the observation model

imate Bayesian filtering techniques. Markovian because past information before the time $t$ is discarded, nonlinear due to functions to deal with and non-Gaussian because of the noisy observations. In this setting, the unobserved signal is denoted $\{\mathbf{X}_t \in \mathcal{X}; t \in \mathbb{N}\}$. This signal has initial distribution $p\left(\mathbf{X}_0\right)$ and dynamical model $p\left(\mathbf{X}_{t+1}|\,\mathbf{X}_t\right)$. The signal has a observable part, denoted as $\{\mathbf{y}_t \in \mathcal{X}; t \in \mathbb{N}\}$, that is also called the current measurement. The measurement process is modeled by the observation distribution $p(\mathbf{X}_{t+1}|\mathbf{y}_{t+1})$.

The main goal in image tracking is to infer (estimate) the filtering distribution $p(\mathbf{X}_t|\mathbf{y}_{1:t})$ (over most recent state) or the posterior distribution $p(\mathbf{X}_{0:t}|\mathbf{y}_{1:t})$ (over a whole trajectory). This inference problem is usually intractable in this type of models that depends on (potentially infinite) time. In probabilistic graphical model theory, exact inference is possible only in specific cases. Among many methods for approximate inference, variational methods [256] and Monte Carlo methods [70] play a major role. In this work. we focus only on Monte Carlo-based methods. Sequential Monte Carlo methods offer a approximate probabilistic framework for recursive dynamic state estimation, that fits with multi-target tracking problem. We refer readers to [70] for a more in-depth treatment of these methods.

When dealing with multiple targets the vector $\mathbf{X}_t$ has to take in account the different targets properly. For this reason, we define the joint state for all target as $\mathbf{X}_t = \{\mathbf{x}_t^1, \mathbf{x}_t^2, \ldots, \mathbf{x}_t^K\}$ and $\mathbf{x}_t^k$ the state of $k$-th target. In single target tracking, $\mathbf{X}_t = \mathbf{x}_t^k$ with $k = 1$. We will omit the index in that case, because the general theory for single-target tracking and multi-target tracking is the same in this description.

The Bayesian rule and the Chapman-Kolmogorov equation enable us to find a sequential formulation of the problem:

$$p(\mathbf{X}_{t+1}|\mathbf{y}_{1:t+1}) \propto p(\mathbf{y}_{t+1}|\mathbf{X}_{t+1}) \int_{\mathbf{X}_{t+1}} p(\mathbf{X}_{t+1}|\mathbf{X}_t)p(\mathbf{X}_t|\mathbf{y}_{1:t})d\mathbf{X}_{t+1}. \tag{2.1}$$

This equation is fully specified by an initial distribution $p(\mathbf{X}_0)$, the dynamical model $p(\mathbf{X}_{t+1}|\mathbf{X}_t)$, and the observation model $p(\mathbf{y}_{t+1}|\mathbf{X}_{t+1})$, as anticipated above. Equation 2.1 is often analytically intractable, because of the integral usually defined in a high dimensional space and the probability distribution are highly nonlinear. Only in few cases admits an analytic solution of problem, for example Gaussian probabilities and linear systems are solved by the Kalman filter. In particle filtering, the filtering distribution at previous time $p(\mathbf{X}_{0:t+1}|\mathbf{y}_{1:t+1})$ is approximated by a set of $N$ weighted particles, *i.e.* $\{(\mathbf{X}_{0:t}^{(n)}, w_t^{(n)})\}_{n=1}^{N}$. Note that the parameter $N$ is usually set by hand, but there exist techniques for estimating the optimal number of particles that minimizes some measure of tracking distortion [184].

The filtering distribution defined as Equation 2.1 can be approximated using the Monte Carlo as:

$$\widetilde{p}(d\mathbf{X}_{t+1}|\mathbf{y}_{1:t+1}) \approx \sum_{n=1}^{N} w_{t+1}^{(n)} \; \delta_{\mathbf{X}_{t+1}^{(n)}}(d\mathbf{X}_{t+1}).$$

where $\mathbf{X}_{t+1}^{(n)}$ has to be sampled from a certain type of distribution called proposal distribution $q(\cdot)$. It turns out that using *importance sampling* the update of the weights is computed according the follows relation (details in [12, 70]):

$$\widetilde{w}_{t+1}^{(n)} = \widetilde{w}_t^{(n)} \; \frac{p(\mathbf{y}_{t+1}|\mathbf{X}_{t+1}^{(n)}) \; p(\mathbf{X}_{t+1}^{(n)}|\mathbf{X}_t^{(n)})}{q(\mathbf{X}_{t+1}^{(n)}|\mathbf{X}_t^{(n)}, \mathbf{y}_{t+1})} \tag{2.2}$$

where $\widetilde{w}_t^{(n)}$ is the un-normalized *importance weight* associated to the $n$-th particle $\mathbf{X}_t^{(n)}$. Note that the importance weights have to be normalized at each time step, that is,

$$w_{t+1}^{(i)} = \frac{\widetilde{w}_{t+1}^{(i)}}{\sum_{j=1}^{N} \widetilde{w}_{t+1}^{(j)}},$$

because Equation 2.1 is defined to hold up to a constant normalization factor. The design of an optimal proposal distribution is a critical task [70]. A common choice is to use the dynamical mode, *i.e.*, $q(\mathbf{X}_{t+1}^{(n)}|\mathbf{X}_t^{(n)}, \mathbf{y}_{t+1}) = p(\mathbf{X}_{t+1}^{(n)}|\mathbf{X}_t^{(n)})$ so that simplifies Equation 2.2 as follows:

$$\widetilde{w}_{t+1}^{(n)} = \widetilde{w}_t^{(n)} \; p(\mathbf{y}_{t+1}|\mathbf{X}_{t+1}^{(n)}). \tag{2.3}$$

The weight at the current time is updated using the weight at the previous time and evaluating the likelihood of the observation $\mathbf{y}_{t+1}$ with respect to the hypothesis $\mathbf{X}_{t+1}^{(n)}$. However, it is possible to construct better proposal distributions, which make use of more recent observations, using object detectors [178], saliency maps [114], optical flow, and approximate filtering methods such as the unscented particle filter. In Sec. 2.2.2, we will analyze a particle filter approach where the proposal is defined with object detectors as in [178].

In practice, the inference algorithm that emerges from this theoretical analysis is very simple. Figure 2.2 reports the particle filtering steps when the importance distribution is equal to the dynamical model. The prior probability distribution $\widetilde{p}(d\mathbf{X}_t|\mathbf{y}_{1:t})$ is represented by a set of weighted hypothesis, called particles. The particles are then sampled according to their weights $w_t^{(n)}$: the higher is the weight, the more particles it generates. Then, a dynamical model $p(\mathbf{X}_{t+1}^{(n)}|\mathbf{X}_t^{(n)})$ is applied to each particle. The observation model $p(\mathbf{y}_{t+1}|\mathbf{X}_{t+1}^{(n)})$ assigns a weight to each

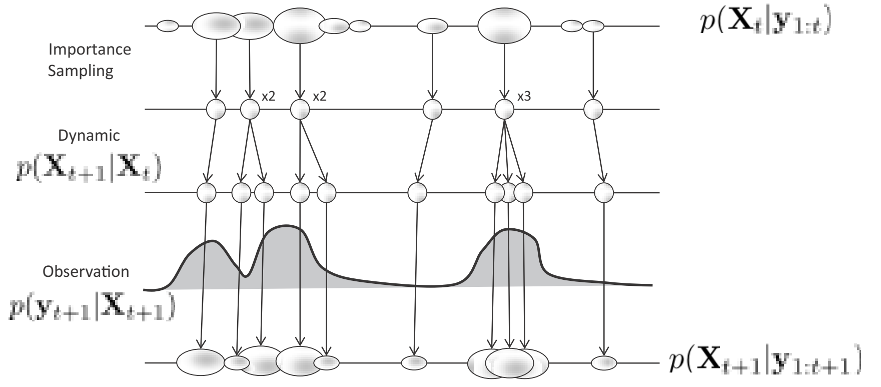

$p(\mathbf{X}_t|\mathbf{y}_{1:t})$
Importance Sampling
x2
x2
x3
Dynamic
$p(\mathbf{X}_{t+1}|\mathbf{X}_t)$
Observation
$p(\mathbf{y}_{t+1}|\mathbf{X}_{t+1})$
$p(\mathbf{X}_{t+1}|\mathbf{y}_{1:t+1})$

**Fig. 2.2.** Particle filtering step. The prior probability distribution is represented by a set of weighted particles. The particles are then sampled according to their weights. And, a dynamical model is applied. The observation model assigns a weight to each particle. Finally, the posterior distribution is approximated by the new set of weighted particles.

particle, that is proportional with the likelihood between the hypothesis and the model. Finally, the filtering distribution $\widetilde{p}(d\mathbf{X}_{t+1}|\mathbf{y}_{1:t+1})$ is approximated by the new set of weighted particles. So, we do not have to explicitly compute the filtering distribution never. This is an huge advantage, because we do not have to enumerate all the possible states and compute their associated probability values; we just compute the probability values for certain states, the ones that are more likely to be interesting to investigate.

When only few particles have considerable weights for the posterior estimate, tracking degenerates to this few particles. This issue is well known in literature as *degeneracy problem* and can be faced introducing a *resampling step* [12, 70]. The resampling step is a selection step used to obtain an "unweighted" approximate empirical distribution $\hat{p}(d\mathbf{X}_{t+1}|\mathbf{y}_{1:t+1})$ of the weighted measure $\tilde{p}(d\mathbf{X}_{t+1}|\mathbf{y}_{1:t+1})$, that is:

$$\hat{p}(d\mathbf{X}_{t+1}|\mathbf{y}_{1:t+1}) \approx \frac{1}{N}\sum_{n=1}^{N} \delta_{\mathbf{X}_{t+1}^{(n)}}(d\mathbf{X}_{t+1}).$$

The basic idea is to discard samples with small weights and multiply those with large weights. The use of a selection step is key to making the sequential Monte Carlo procedure effective; see [70] for details on how to implement this black box routine.

To summarize, we report the general procedure in Algorithm 1. Note that in literature the target tracking formulation has been defined as both the filtering distribution $p(\mathbf{X}_{t+1}|\mathbf{y}_{1:t+1})$ (over the most recent state) and the posterior distribution $p(\mathbf{X}_{0:t+1}|\mathbf{y}_{1:t+1})$ (over the whole trajectory). In the latter case, the theory and the algorithm is similar: We just replace $\mathbf{X}_t$ with $\mathbf{X}_{0:t}$, where it appears in the theory above and in Algorithm 1. In this chapter, we will use the formulation with the filtering distribution.

**Algorithm 1**: Particle filtering algorithm based on importance sampling.

```
1. Initialization
   for $i = 1$ to $N$ do
      $\mathbf{X}_0^{(i)} \sim p(\mathbf{X}_0)$

for $t = 1 \ldots$ do
   2. Importance sampling
      for $i = 1$ to $N$ do {Predict the next state}
         $\widetilde{\mathbf{X}}_{t+1}^{(i)} \sim q\left(d\mathbf{X}_{t+1}^{(i)} \middle| \widetilde{\mathbf{X}}_t^{(i)}, \mathbf{y}_{1:t},\right)$
         $\widetilde{\mathbf{X}}_{0:t+1}^{(i)} \leftarrow \left(\mathbf{X}_{0:t}^{(i)}, \widetilde{\mathbf{X}}_{t+1}^{(i)}\right)$
      for $i = 1$ to $N$ do {Evaluate the importance weights}
         $\widetilde{w}_{t+1}^{(i)} \leftarrow \frac{p\left(\mathbf{y}_t|\widetilde{\mathbf{X}}_{t+1}^{(i)}\right) p\left(\widetilde{\mathbf{X}}_{t+1}^{(i)}|\widetilde{\mathbf{X}}_t^{(i)}\right)}{q\left(\widetilde{\mathbf{X}}_{t+1}^{(i)} \middle| \widetilde{\mathbf{X}}_t^{(i)}, \mathbf{y}_{1:t+1}\right)}$
      for $i = 1$ to $N$ do {Normalize the importance weights}
         $w_{t+1}^{(i)} \leftarrow \frac{\widetilde{w}_{t+1}^{(i)}}{\sum_{j=1}^N \widetilde{w}_{t+1}^{(j)}}$
   3. Selection
      Resample with replacement $N$ particles $\left(\mathbf{X}_{t+1}^{(i)}; i = 1, \ldots, N\right)$ from the set
      $\left(\widetilde{\mathbf{X}}_{t+1}^{(i)}; i = 1, \ldots, N\right)$ according to the normalized importance weights $w_{t+1}^{(i)}$
```

In the next two sections, we first describe an approach that explicitly deal with multi-target tracking (Sec. 2.2.1) by making an approximation instead to treat the problem as detailed in this section. Second, a tracking-by-detection algorithm procedure will be discussed (Sec. 2.2.2).

### 2.2.1 Hybrid Joint-Separable Filter

The Hybrid Joint-Separable (HJS) filter [142] is an instance of particle filter that represents a theoretical grounded compromise between dealing with a strict joint process [113] and instantiating a single independent tracking filter for each distinct object. Roughly speaking, HJS alternates a separate modeling during the sampling step with a joint formulation using a hybrid particle set in the dynamical and observational steps. In this section, we briefly discuss the main characteristics of HJS filter, we refer to [142] for more details.

The rule that permits the crossing over joint-separable treatments is based on the following approximation:

$$p(\mathbf{X}_t|\mathbf{y}_{1:\mathcal{T}}) \approx \prod_k p(\mathbf{x}_t^k|\mathbf{y}_{1:\mathcal{T}}) \tag{2.4}$$

that is, the joint posterior could be approximated via product of its marginal components ($k$ indexes the objects). This assumption enables us to sample the particles in each state space independently (thus requiring a linear proportionality between the number of objects and the number of samples), and to update the weights in the joint state space in order to model interactions between targets.

The updating exploits a joint dynamical model which builds the distribution $p(\mathbf{X}_{t+1}|\mathbf{X}_t)$ (explaining how the system does evolve) and a joint observational model that provides estimates for the distribution $p(\mathbf{y}_{t+1}|\mathbf{X}_{t+1})$ (explaining how the observations are related to the state of the system). Both the models take into account the interactions among objects; in particular the joint dynamical model accounts for physical interactions between the targets, thus avoiding track coalescence of spatially close targets. The joint observational model quantifies the likelihood of the single measure $\mathbf{y}_{t+1}$ given the state $\mathbf{X}_{t+1}$, considering inter-objects occlusions.

Given the joint formulation, we can easily derive the independent dynamics and observation models of each object marginalizing out specific variables as follows:

$$p(\mathbf{x}_{t+1}^k|\mathbf{x}_t^k) = \int_{\mathbf{X}_{t:t+1}^{\neg k}} p(\mathbf{X}_{t+1}|\mathbf{X}_t)p(\mathbf{X}_t^{\neg k}|\mathbf{y}_{1:t})d\mathbf{X}_{t:t+1}^{\neg k} \tag{2.5}$$

$$p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}^k) = \int_{\mathbf{X}_{t+1}^{\neg k}} p(\mathbf{y}_{t+1}|\mathbf{X}_{t+1})p(\mathbf{X}_{t+1}^{\neg k}|\mathbf{y}_{1:t})d\mathbf{X}_{t+1}^{\neg k} \tag{2.6}$$

where the superscript $^{\neg k}$ addresses all the targets but the $k$th. Eq. 2.6 and 2.5 are usually intractable, but it is possible to estimate the posterior distribution using HJS without explicitly compute them.

The joint dynamical model is approximated in the following way:

$$p(\mathbf{X}_{t+1}|\mathbf{X}_t) = p(\mathbf{X}_{t+1}) \prod_{k=1}^{K} q(\mathbf{x}_{t+1}^k|\mathbf{x}_t^k) \tag{2.7}$$

where $q(\mathbf{x}_{t+1}^k|\mathbf{x}_t^k)$ is the single target dynamical model, that spread independently the particles of each target, and $p(\mathbf{X}_{t+1})$ is a joint factor that models the interaction among the targets. The model, through the prior $p(\mathbf{X}_{t+1})$, avoids that multiple targets with single motion described by $q(\mathbf{x}_{t+1}^k|\mathbf{x}_t^k)$ collapse in the same location.

The particle filter dynamic process is split in two step: 1) apply the dynamic of the single target hypotheses, and 2) jointly evaluate the interactions among the hypotheses of all the targets. In particular, we model $q(\mathbf{x}_{t+1}^k|\mathbf{x}_t^k) = \mathcal{N}(\mathbf{x}_{t+1}^k; \mathbf{x}_t^k, \Sigma)$, that is, a first order autoregressive model with Gaussian noise. The joint factor $p(\mathbf{X}_{t+1})$ can be viewed as an exclusion principle: two or more targets cannot occupy the same volume at the same time. In [142], $p(\mathbf{X}_{t+1})$ is modeled with a pairwise Markov Random Field (MRF). Inference on the MRF is performed with belief propagation. The effect is that hypotheses that do not agree with the exclusion principle will have a low probability to exist.

The joint observational model relies on the representation of the targets, that here are constrained to be human beings. Human representation assumes the human body in three parts [113]: head, torso, and legs. For the sake of clarity, we assume the body as a whole volumetric entity, described by its position in the

3D plane, with a given volume and appearance. The joint observational model is defined as standard template matching method. The idea is to correlate a template that is usually captured at the first frame with the whole image or a set of hypothesis. In particle filtering, it evaluates the distance between the histograms of the template and the hypotheses. HJS additionally involves also a joint reasoning captured by an *occlusion map*. The occlusion map is a 2D projection of the 3D scene which focuses on the particular object under analysis, giving insight on what are the expected visible portions of that object. This is obtained by exploiting the hybrid particles set $\{\mathbf{x}_{t+1}^{(p)}\}_{p=1}^{NK}$ in an incremental visit procedure on the image plane[2]: the hypothesis nearest to the camera is evaluated first, its presence determines an occluding cone in the scene, where the confidence of the occlusion depends on the observational likelihood achieved. Particles farther in the scene which fall in the cone of occlusion of other particles are less considered in their observational likelihood computation. The process of map building is iterated as far as the farthest particle in the scene and the observation model is defined as:

$$p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}^{(p)}) \propto e^{-\lambda\left(\mathrm{fc}(\mathbf{x}_{t+1}^{(p)},\mathbf{y}_{t+1})+\mathrm{bc}(\mathbf{x}_{t+1}^{(p)},\mathbf{y}_{t+1})\right)} \tag{2.8}$$

where $\mathrm{fc}_p$ is the foreground term, *i.e.*, the likelihood that an object matches the model considering the un-occluded parts, and $\mathrm{bc}_p$, the background term, accounts for the occluded parts of an object. These terms are computed accounting for the occlusion map, that deals with the partial occlusions among different persons on the image plane. The parameter $\lambda$ tunes the variance of the distribution, in fact sometimes it is defined as $\lambda = \frac{1}{2\,\sigma^2}$ for a better tuning, where $\sigma$ has a similar interpretation of the variance of the Gaussian dstribution.

### 2.2.2 Tracking-by-detection

We discussed in Section 2.1 many tracking-by-detection strategies that have been proposed in literature. In this section, we present the technique proposed in [50, 178], because it seems one of the most mathematical sound: It nicely fits the particle filtering framework presented in the previous section. This technique is used in an algorithm later in the thesis (Section 4.5).

The idea first proposed in [178] is very simple yet powerful. Analyzing Equation 2.2 and its approximation of Equation 2.3, it turns out that most of the particle filter-based algorithms presented in literature – before the seminal work in [178] – discard a very important information in the proposal distribution $q(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{t+1})$, that is, the current measurement $\mathbf{y}_{t+1}$. This is equivalent to state: $q(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{t+1}) = q(\mathbf{X}_{t+1}|\mathbf{X}_{0:t})$. However, the measurement will play a fundamental role in deciding in which regions of the state space to generate (sample) new hypotheses. Imagine that you want to drive to a certain destination. The sensors are your eyes and your basic knowledge of the neighborhood. You observe and predict the next action to do (turn right, turn left or go straight). Not using $\mathbf{y}_{t+1}$ in the proposal is like having a satellite navigator that detects your position and the goal position (not the map) and not use it. Instead, using it would rule

[2] Note that calibration parameters are necessary to define a ordering between particles and thus to built the occlusion map.

out a lot of possible wrong actions and directions and keep low the number of configurations to explore in the state space.

In [178] the authors fully understand and take advantage from this "full" proposal distribution (with measurements). In particular, they define it as follows:

$$q(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{t+1}) = \alpha\, q_d(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{t+1}) + (1-\alpha)\, q(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}) \tag{2.9}$$

where $q(\mathbf{X}_{t+1}|\mathbf{X}_{0:t})$ is the standard dynamical model, $q_d(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{t+1})$ is a distribution that samples hypotheses where an object detector gives positive responses and $\alpha$ is a weight that bias the two terms. Note that in this case we just have two proposal, but the equation can be generalized as:

$$q(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{t}) = \sum_{m=1}^{M} \alpha_m\, q_m(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{t+1}) \tag{2.10}$$

where we have $M$ proposals $q_m(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{t+1})$ that evaluate multiple measurements and $\alpha_m$ is a parameter that as to satisfies $\sum_{m=1}^{M} \alpha_m$. Sampling from the mixture of proposals is easy: The idea is to sample $N \cdot \alpha_m$ particles from each proposal. The only drawback is that the number of particles associated to each proposal has to be high enough to obtain a robust statistics.

The main characteristics of this strategy are that:

- the generated hypotheses are more consistent with the measurements,
- the idea is extended to multiple proposal distributions that evaluate different measurements,
- sampling is easy if it is possible to sample from the components of the mixture,
- it works well in practice.

We will use this method later in the next chapters; we introduced it here, because it is simpler to understand when the formulation of particle filtering is presented.

## 2.3 On-line Feature Selection for Partial Occlusions

Most of the state-of-the-art tracking algorithms are prone to error when dealing with occlusions, especially when the involved targets are hardly discernible in appearance. We propose a mechanism of *online subjective feature selection* embedded into a tracker, that selects and employs the most discriminant features characterizing a single target with respect to the neighboring targets. Our approach takes inspiration from [59], consisting in a feature selection policy for discriminating effectively foreground (the moving objects) and background (the static scene) in a video surveillance context. The improvement here consists in devising a feature selection technique which operates among different foreground objects, whose number and appearance may change over time. Moreover, attention is devoted to maintain the computational effort limited while still achieving performances higher than those of the original framework. Therefore, the feature selection process is activated only in case of proximity among objects. When an occlusion occurs, the mechanism is frozen and only the previously selected features are used into the observational step.

In this section, we extend the joint observational model of the HJS framework (Sec. 2.2.1). But, we would like to stress the fact that the contribute is easily generalizable to whatever multi-target particle filter, in which the observations are evaluated in a joint space (*i.e.*, taking into account dependencies among all the tracked objects). The extension translates in a new term in the observational model, the *foreground feature discrimination term* ff:

$$p(\mathbf{y}_{t+1}|\mathbf{x}^{(p)}_{t+1}) \propto e^{-\lambda\left(\mathrm{fc}(\mathbf{x}^{(p)}_{t+1},\mathbf{y}_{t+1})+\mathrm{bc}(\mathbf{x}^{(p)}_{t+1},\mathbf{y}_{t+1})\right)} e^{-\lambda_{\mathrm{ff}}\ \mathrm{ff}(\mathbf{x}^{(p)}_{t+1},\mathbf{y}_{t+1})} \tag{2.11}$$

where the first exponential term is the same of Equation 2.8. The foreground feature discrimination term ff is introduced in occlusion cases in order to help appearance disambiguation among similar targets which stand spatially close, finding the most discriminative parts of an object with respect to the other surrounding objects.

In the following, we explain how this term is evaluated in a two-objects scenario, generalizing then to the case with more objects.

### 2.3.1 Two-objects case

The stream of data for the two-object case is depicted in Figure 2.3. The first step is to choose a set of candidate features. Here, we use a small set of $M$ features based on RGB color histogram, which it has been shown in [59] to be experimentally appropriate for tracking applications. The feature set contains the linear combination of R, G, B pixel values: $\mathcal{F} = \{w_1 R + w_2 G + w_3 B \mid w_* \in [-2, -1, 0, 1, 2]\}$, pruning out redundant combinations. Such class of features is computational fast to manage, and shows adequate expressiveness. Then, $M$ histograms of features ($b$ bins) have been built considering each of the two objects' appearances (first phase in Figure 2.3).

Second, the histograms of features are combined together to distill a combined feature, tuned to discriminate between the two objects in the current frame. In particular, the log-likelihood ratio has been computed as

$$L = \log \frac{p}{q},$$

where $p$ and $q$ are the histograms of a single feature for the first and second object, respectively (second phase in Figure 2.3). Log-likelihood is naturally discriminative: thresholding $L$ at zero is equivalent to use a maximum likelihood rule to classify the two objects. This feature permits thus to rewrite the possible multimodal distributions $p$ and $q$ into a unimodal distribution. Finally, we introduce an evaluation criterion which measures the separability that feature $L$ induces between the two classes. Likewise [59], we employ the two-class variance ratio, which, given two class distributions $q$ and $p$ (their histograms), is defined as:

$$\mathrm{VR}(L;p,q) = \frac{\mathrm{var}(L;(p+q)/2)}{\mathrm{var}(L;p)+\mathrm{var}(L;q)}. \tag{2.12}$$

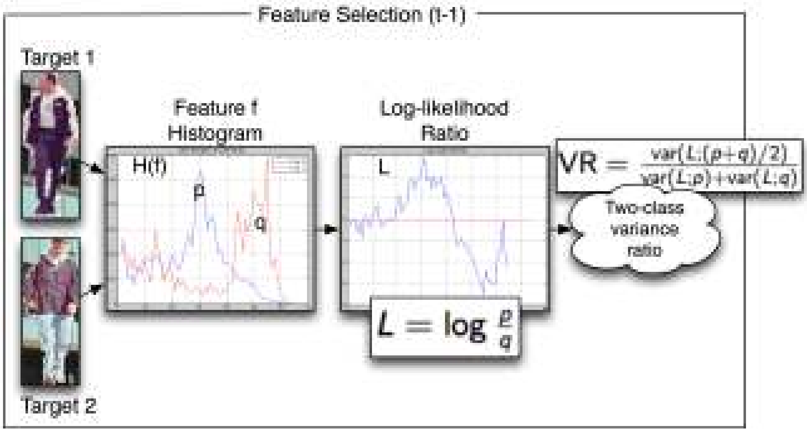


**Fig. 2.3.** Feature selection analysis of the two-object case. Features are extracted and compared using the log-likelihood ratio. Then, the variance ratio scores how discriminant is each feature for that case.

The denominator enforces that the within-class variances should be small for both objects' classes, while the numerator rewards cases where values associated with two different objects are widely separated. At the end of the process we have, for each moving object, a new feature set $\mathcal{F}_s = \{f_1, \ldots, f_s\} \subseteq \mathcal{F}$, built by selecting the the top $N_f$ most discriminative individual features (ordered by decreasing VR). For each frame of the video we compute the set $\mathcal{F}_s$ only if the two objects are very close.

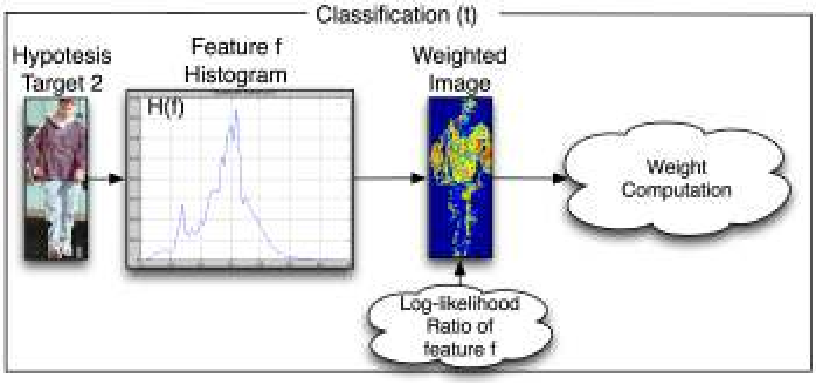
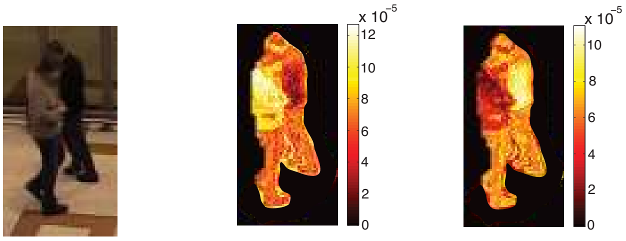


**Fig. 2.4.** Process to extract the discrimination map and the foreground feature discrimination term.

In [59], the feature selection method is embedded in a mean-shift tracking system. Here, our method uses the selected features $\mathcal{F}_s$ in order to build a map for each object $k$ involved in an occlusion, called *discrimination map* $F_k$, that favors the pixels corresponding to the discriminative parts of an object (Figure 2.4). Such map is obtained fusing the rank of the discriminative features as a weighted sum, *i.e.*,

$$F_k = \sum_{s=1}^{N_f} \mathrm{VR}_s \; g(L_s, I) \tag{2.13}$$

where $s$ indexes the log-likelihood ratio features $\{L\}$, and $g$ is the function that maps the 2D rendering $I$ of a person to the discrimination map, assigning the values of $L_s$ to the image $I$. An example of discrimination maps are shown in Figure 2.5. In order to avoid that the background clutter distracts the feature selection mechanism, we remove it using the background subtraction algorithm in [232]. Note that in Figure 2.5 the background is ruled out.



(a) Image (b) Object 1 (c) Object 2

**Fig. 2.5.** Discrimination maps for two objects during an occlusion; brighter pixels are those whose values ensure higher discrimination.

Feature selection is stopped when an occlusion occurs, because the algorithm risks to learn discriminant pixels for the nearest targets to the camera, that is wrong for an occluded target. A discrimination map is built for each sample hypothesis, using the feature set $\mathcal{F}_s$ selected at previous time, and employed to assign a reasonable weight in the observational step. Given a particle $\mathbf{x}_t^{(p)}$, the foreground feature discrimination term ff is computed using the discrimination map $F_k$ as follows:

$$\mathrm{ff} = 1 - p(\mathbf{y}_t | \mathcal{F}_s, \mathbf{x}_t^{(p)}) = 1 - \sum_u F_k(u) \; \delta(B_t(u)) \tag{2.14}$$

where $B_t$ is a binary map given from the background subtraction algorithm [232], $\delta$ is the Kronecker delta and $u$ indexes the pixels. This term will be higher for the particles far from the discriminant parts of the object and vice versa. When exponentiating it in Eq. 2.11, we will have high probability for the discriminant parts and low probability for the other parts.

### 2.3.2 Multi-objects case

Our goal is to select the features that discriminate between a particular object $h$ and the surrounding objects. We decompose such task as that of finding a set of ranked features for a single pairwise discrimination (the previous section), repeating the process for all the couple of objects that include $h$. In particular, assuming that the discriminative features for an object with respect to the surrounding single objects are given, we can group the set $\{\mathcal{F}_s^{h,1}, \dots, \mathcal{F}_s^{h,K}\}$, where the $\mathcal{F}_s^{h,k}$ is the set of discriminative features related to object $h$ with respect to the object $k$. The features for the multiple discrimination are retrieved by exploiting an intersection operation:

$$\mathcal{F}_{\cap}^{h} = \bigcap_{k=1}^{K} \mathcal{F}_s^{h,k}. \tag{2.15}$$

The new feature set $\mathcal{F}_{\cap}^{h}$ contains the common features of every single set $\mathcal{F}_s^{h,k}$, that is the set of the best discriminative features for the object $h$ as compared to all the other surrounding entities.

## 2.4 Experiments

In this section, we first compare the HJS filter with a state-of-the-art filter for tracking, that is, the Multi-Hypothesis Kalman Filter (MHKF). We prove that particle filter-based approaches outperform Kalman-based methods in terms of accuracy and number of tracks generated. Then, we analyze the on-line feature selection method for partial occlusions proposed in Sec. 2.3. The experiments on HJS show that the results are improved by using the feature selection mechanism.

### 2.4.1 Evaluation

Let's first analyze which evaluation metrics and the available public datasets we can use for tracking. As for the quantitative evaluation, we used the standard measurements presented in [228], that consist of:

- False Positives (**FP**): An estimate exists that is not associated with a ground truth object;
- Multiple Objects (**MO**): Two or more ground truth objects are associated with the same estimate;
- False Negatives (**FN**): A ground truth object exists that is not associated with an estimate;
- Multiple Trackers (**MT**): Two or more estimates are associated with the same ground truth.

The single-frame values are averaged over time for estimating an overall statistics. In addition, we also provide an evaluation in terms of:

- Tracking Success Rate (**TSR**) summarizes the overall tracker accuracy over time;
- Mean Error (**ME**) with respect the ground truth.

| | FP | MO | FN | MT | TSR |
|---|---|---|---|---|---|
| MHKF | 0.279 | 0.009 | **0.203** | 0.212 | 0.624 |
| HJS | **0.086** | **0.007** | 0.279 | **0.042** | **0.712** |

**Table 2.1.** Tracking results comparison on PETS 2009 dataset: task S2, video L1, View001.

Other evaluation measurements can be used. It is worth to mention the recent one proposed in [126]; we used them in Sec. 6.5.1.

Some of the most challenging datasets that are publicly available are presented at the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS). In particular, we take in consideration the datasets PETS 2006 [4], PETS 2007 [5] and PETS 2009 [6]. PETS 2006 contains seven datasets, recorded at Victoria Station in London, UK. In PETS 2007, nine datasets were recorded at BAA Glasgow Airport. PETS 2009 involves up to approximately forty actors and considers crowd image analysis and include crowd count and density estimation, tracking of individual(s) within a crowd, and detection of separate flows and specific crowd events. The resolution of all sequences are PAL standard ($768 \times 576$ pixels, 25 fps) and compressed as JPEG image sequences. All sequences contains the calibration data that are useful for HJS.

### 2.4.2 HJS vs. MHKF

The testing session has been focused on the popular publicly available dataset PETS 2009 [6], in order to compare the performances between HJS filter and MHKF on challenging real world data. In particular, we choose S2 dataset, using the first camera view (“View001”) of the first sequence because we took part in PETS 2009 evaluation. The objective was to track all of the persons in the sequence with both methods in a monocular setup. In order to give an evaluation and validation of the methods, the sequence has been manually labeled generating the ground truth of the targets. The ground truth consists of: the identifier (ID) of each target, its 3D position on the ground plane, and its 2D bounding box which is associated to a specific view.

We used the online background subtraction method presented in [27] for both the algorithms. The background modeling method initialization step needs 220 frames to create the first background model, thus people tracking starts from the frame number 220 to the end of the sequence (795). In Fig. 2.6, we can appreciate qualitatively the results of the background subtraction algorithm.

The tracking results, averaged over all the frames of the sequence and for all the moving targets, are summarized in Table 2.1. HJS filter performs better than MHKF considering FP, MO, MT, and TSR, because MHKF generates multiple tracks for a single target at each time. On the other had, the FN ratio is higher for HJS, because sometimes the tracking of an target is lost and it converges toward another target or the clutter. The TSR gives us a general value of the tracking reliability and it summarizes the performances. It is clear from the results that HJS is more reliable than MHKF.

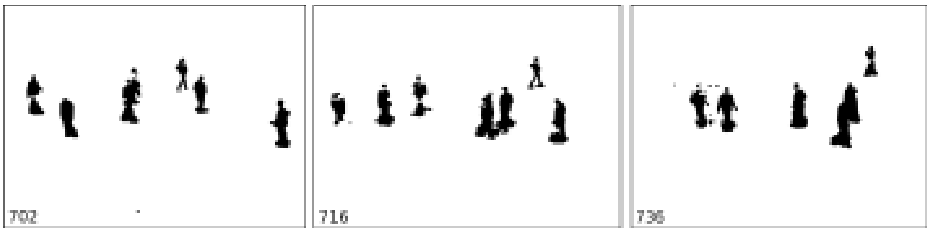


**Fig. 2.6.** Foreground images (moving objects in black) for frames 702, 716, and 736 of task S2, video L1, View001.

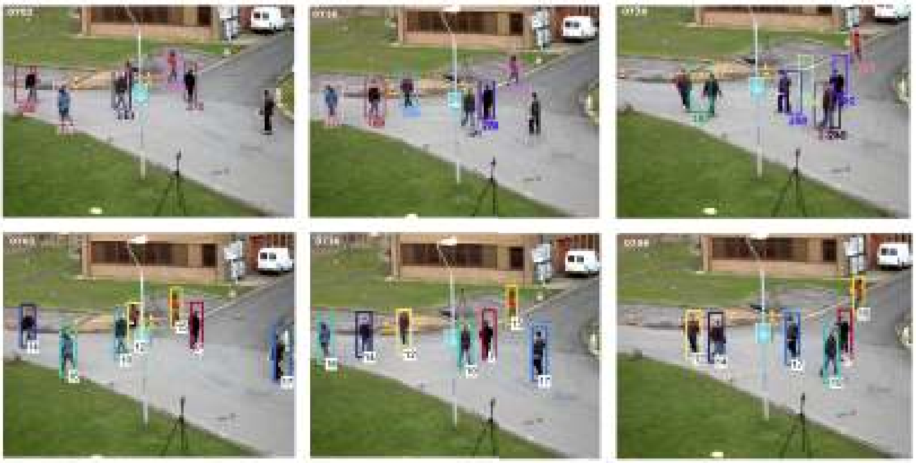


**Fig. 2.7.** Tracking results for frames 702, 716, and 736 of task S2, video L1, View001: first row shows MHKF and second row show HJS filter.

In Fig. 2.7, we report some tracking results for a qualitative evaluation. In particular, we show three frames of the task S2, video L1, View001 from PETS 2009 dataset. First row shows the tracking results regarding MHKF, whereas second row shows HJS results. From the experiments we find that drawbacks of MHKF are: 1) targets are tracked with multiple tracks, leading to a proliferation in the number of tracks; 2) after an occlusion the target ID changes, *i.e.*, a new track instance is created; 3) MHKF tracking fails when people motion is non-linear. Instead, HJS overcomes the problems of MHKF: 1) only one track is kept for each target, because the data association is inherent in the particle filtering formulation; 2) after an occlusion the target ID is kept, thanks to the integration of the occlusion map into the observation model; 3) HJS can deal with non-linearity characterizing people motion. However, also HJS has some problems, in particular when complete occlusions occur. In that case, the tracker cannot infer the position of the target because of the lack of foreground observations for the occluded target. Thus, tracking of a target fails in case of long-term occlusions.

Integrating MHKF and HJS can lead to reduce erroneous track associations. HJS can help MHKF in merging redundant tracks belonging to the same target, re-

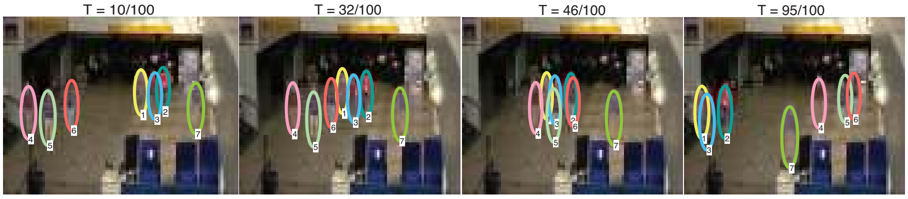


**Fig. 2.8.** Synthetic video example tracked by our method. Occlusions occur among different objects are correctly handled.

| | ME | FP | MO | FN | MT | TSR |
|---|---|---|---|---|---|---|
| HJS | 0.64 | 0.12 | 0 | 0.12 | 0 | 0.87 |
| HJS + FS | **0.53** | **0.09** | 0 | **0.09** | 0 | **0.91** |

**Table 2.2.** Results comparison on synthetic videos.

ducing the problem of track proliferation. MHKF can help HJS in re-initialization after a track loss and in deleting track not corresponding to foreground observations.

### 2.4.3 HJS with Online Feature Selection

The testing session of the proposed online feature selection mechanism has been focused on HJS with three datasets: a synthetic dataset and two public and challenging datasets (PETS 2006 and 2007). As comparative tracking framework, we consider the original version of the HJS filter proposed in [142]. In a wider sense, such comparison is highly valuable, being HJS a tracker with strong performances.

Since PETS 2006 and 2007 sequences do not contain the ground truth of the position of the objects, as preliminary test and quantitative analysis we create a set of *synthetic* videos. In particular, 8 synthetic sequences (of 100 frames each) have been built by superimposing different static pedestrian images on a static background, mimicking the 3D scenario by applying scaling on the silhouette. The synthetic pedestrians move in the scene, with a dynamics similar to the that of the PETS videos. The comparison of our approach with HJS on this synthetic dataset (see Fig. 2.8) has been performed using a different number of moving objects in the range $[2, 7]$. The tracking results, averaged over all the experiments and for all the moving objects, are summarized in Table 2.2 (the lower the better, except TSR): it can be noted that the proposed approach outperforms HJS, especially in terms of ME and TSR.

Another interesting point is to understand the behavior of the proposed approach while increasing the number of moving objects involved in a occlusion. This enables us to find the maximum number of targets that the feature selection algorithm can deal without decreasing the performances. Such analysis is presented in Table 2.3, for the ME index. The plot clearly shows that finding the discriminative features with a high number of persons is a hard task – the proposed approach is effective in occlusion cases for group of at most 5 people. Note that we have a gating function that limits the surrounding objects to consider in the online feature selection mechanism. Thus, for each person the algorithm always works better if

| Num. of Objs | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| HJS | 0.55 | 0.96 | 0.48 | 0.55 | **0.66** | **0.55** |
| HJS + FS | **0.37** | **0.53** | **0.48** | **0.54** | 0.69 | 0.66 |

**Table 2.3.** ME increasing the number of people.

he/she has up to 4 surrounding people. This is a reasonable assumption because people usually walk in small groups. We will again see this concept in the following Chapters.

Concerning the real dataset, qualitative evaluations on PETS dataset have been carried out exploiting different videos of varying length. The achieved results mirror those gathered on the synthetic trials. Actually, our feature selection strategy provides tracking performances that qualitatively and in average are comparable to those obtained by HJS, outperforming the latter in the case of occlusions. Two examples are shown in Fig. 2.9(a) and 2.9(b).

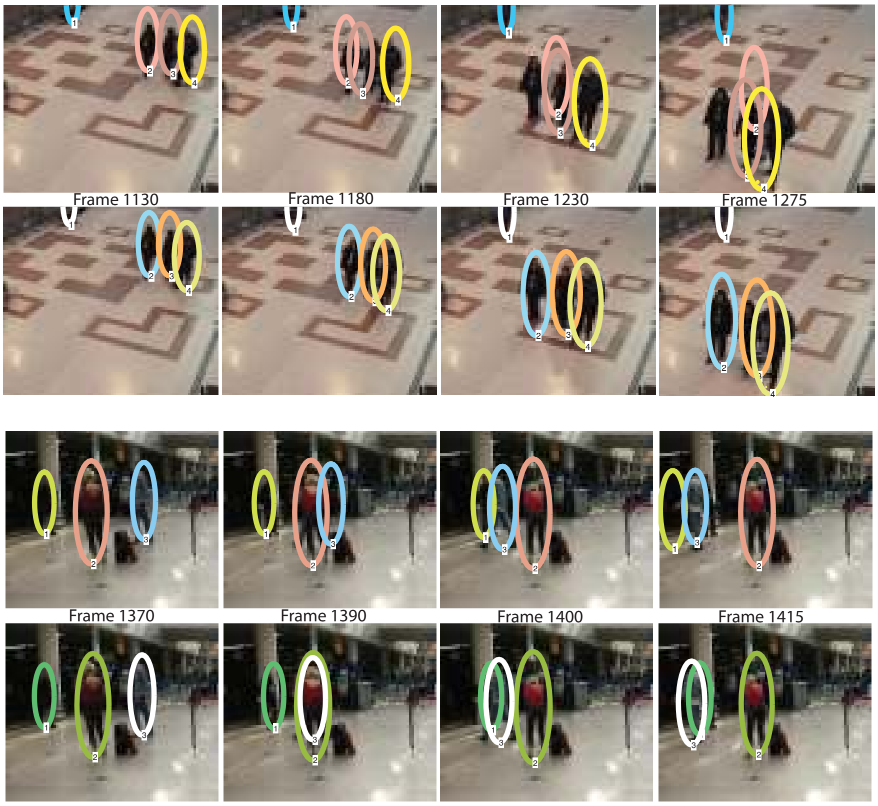


**Fig. 2.9.** A comparison of HJS (first and third row) and our method (second and forth row): sequence S5-T1-G from PETS 2006 (first two rows) and sequence S07 from PETS 2007 (last two rows).

## 2.5 Conclusions

In this chapter, we first analyzed the state of the art of target tracking from theoretical and practical perspectives, and the several issues that are involve. Then, we discussed the theory behind multi-target tracking using particle filtering to perform approximated inference in a state-space model.

Two methods for multi-target tracking have been compared using a recent challenging dataset: the multiple-hypothesis Kalman filter and the HJS filter. The results showed the robustness of both approaches. In particular, HJS performs better than MHKF when occlusions occur while keeping the identity of the target after occlusion whereas MHKF tends to generate a new target ID. However, when dealing with similar targets HJS is prone to error with respect to MHKF. For this reason, a mechanism of *online subjective feature selection* has been proposed. The idea to distill a pool of features discriminating one object with respect to the surrounding ones, so permitting to deal with occlusions among multiple persons. The results show an increase in accuracy when embedding this strategy in a particle filter. It is worth noting that any particle filter (not only HJS) in the joint state space can exploit the advantage of the proposed model.

# 3

# Attentional Policies for Tracking

In the previous chapter, a lot of attention has been focused on the design of a tracking system in order to maximize the prediction accuracy, the time and space performances. These are very important parameters when looking at the tracking problem from the engineering point of view. In this chapter, we want to renew a question that actually has already been discussed a lot by the computer vision community. The question is: Should we emulate the human system or at least take inspiration from it to build our tracking algorithms (or even other computer vision methods)? The question is about two strategies to deal with a computer vision problem: the engineering way and the perceptual way. The former is a class of solutions aimed at solving specific problems to reach the state-of-the-art results. The latter instead takes inspiration from how the human perceptual system works and trying also to understand how the brain works.

In the previous chapter, we have already seen some examples of the former, and we will see them extensively in the next chapters. In this chapter, we want to investigate the perceptual point of view to create new perspectives for tracking. We do not want to decide here which solution is better, but we want to highlight that both need to exist. Indeed, in this chapter we are going to deal with the tracking problem taking inspiration by how humans track and recognize objects. Let's see how it is possible.

## 3.1 Introduction

Humans track and recognize objects effortlessly and efficiently, exploiting attentional mechanisms [60, 207] to cope with the vast stream of data continuously acquired. In this chapter, we exploit the human visual system as inspiration to build a system for simultaneous object tracking and recognition from gaze data. An attentional strategy is learned online to choose fixation points which lead to low uncertainty in the location of the target object. Our tracking system is characterized by of two interacting pathways, being this separation of responsibility a common feature in models from the computational neuroscience literature. It is in fact believed to reflect a separation of information processing into ventral and dorsal pathways in the human brain [179].

The *identity* pathway (ventral) is responsible for comparing observations of the scene to an object template using an appearance model, and on a higher level, for classifying the target object. The identity pathway consists of a two hidden layer deep network. The top layer corresponds to a multi-fixation Restricted Boltzmann Machine (RBM) [143], as shown in Figure 3.1, which accumulates information from the first hidden layers at consecutive time steps. For the first layers, we use (factored)-RBMs [105, 202, 238, 263], but autoencoders [249], sparse coding [127, 180], two-layer ICA [134] and convolutional architectures [146] could also be adopted.

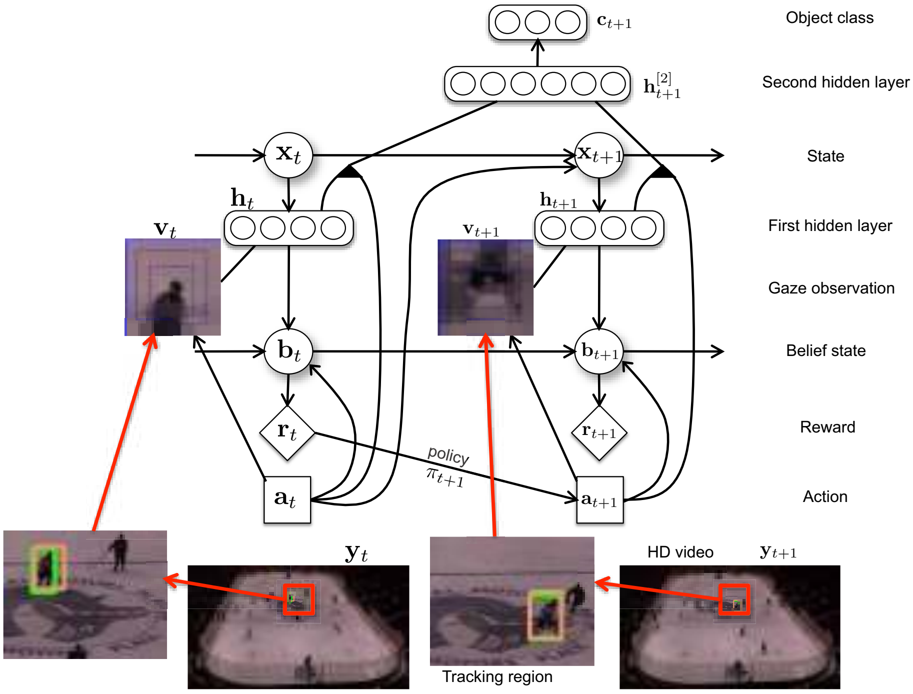


**Fig. 3.1.** From a sequence of gazes $(\mathbf{v}_t, \mathbf{v}_{t+1}, \ldots)$, the model infers the hidden features $\mathbf{h}$ for each gaze (that is, the activation intensity of each receptive field), the hidden features for the fusion of the sequence of gazes and the object class $\mathbf{c}$. The location, size, speed and orientation of the tracking region are encoded in the state $\mathbf{x}_t$. The actions $\mathbf{a}_{t+1}$ follow a learned policy $\pi_{t+1}$ that depends on the past rewards $\{r_1, \ldots, r_t\}$. The reward is a function of the belief state $\mathbf{b}_{t+1} = p(\mathbf{x}_{t+1}|\mathbf{a}_{1:t+1}, \mathbf{h}_{1:t+1})$, also known as the filtering distribution. Unlike typical partially observed Markov decision models (POMDPs), the reward is a function of the beliefs. In this sense, the problem is closer to one of sequential experimental design. With more layers in the ventral $\mathbf{v} - \mathbf{h} - \mathbf{h}^{[2]} - \mathbf{c}$ pathway, other rewards and policies could be designed to implement higher-level attentional strategies.

The *control* pathway (dorsal) is responsible for aligning the object template with the full scene, so the remaining modules can operate independently of the object's position and scale. This pathway is separated into a localization module and a fixation module which work cooperatively to accomplish this goal. The localization module is implemented as a particle filter [69] which estimates the location, velocity and scale of the target object. We make no attempt to implement such states with neural architectures, but it seems clear that they could be encoded with grid cells [170] and retinotopic maps as in V1 and the superior colliculus [97, 212]. The fixation module learns an attentional strategy to select fixation points relative to the object template. These fixation points are the centers of partial template observations, and are compared with observations of the corresponding locations in the scene using the appearance model (see Figure 3.2). Reward is assigned to each fixation based on the uncertainty in the estimate of the target location at each time step. Note that different utilities can be used to reach different goals. For example, in [176] various measures of uncertainty are presented for the decision making in a Bayesian optimal-experimental design: probability gain, Shannon entropy, and Kullback-Leibler distance. In this work, the fixation module uses the reward signal to adapt its gaze selection policy to achieve good localization.

The proposed system can be motivated from many different perspectives. First, as we have seen in the previous chapter many particle filters have been proposed for image tracking, but these typically use simple observation models such as B-splines [111] and color templates [178]. Instead, here we use RBMs that are able to automatically learn features from images. RBMs are more expressive models of shape, and hence we conjecture that they will play a useful role where simple appearance models fail. Second, from a deep learning computational perspective, this work allows us to tackle large images and video, which is typically not possible due to the number of parameters required to represent large images in deep models. The use of fixations synchronized with information about the state (e.g. location and scale) of such fixations eliminates the need to model the entire frame. Third, the system is invariant to image transformations encoded in the state, such as location, scale and orientation. This kind of invariance it is hard to obtain using only the RBMs themselves. Fourth, from a dynamic sensor network perspective, this chapter presents a very simple, but efficient and novel, way of deciding how to gather measurements dynamically. Lastly, in the context of psychology, the proposed model realizes to some extent the functional architecture for dynamic scene representation of [207]. The rate at which different attentional mechanisms develop in newborns (including alertness, saccades and smooth pursuit, attention to object features and high-level task driven attention) guided the design of the proposed model and was a great source of inspiration [60].

## 3.2 Related Work

Our attentional model can be seen as building a saliency map [132] over the target template. Previous work on saliency modeling has focused on identifying salient points in an image using a bottom up process which looks for outliers under some local feature model (which may include top down information in the form of a task dependent prior, global scene features, or various other heuristics). These features can be computed from static images [24, 240], or from local regions of space-time [90] for video. Additionally, a wide variety of different feature types have been applied to this problem, including engineered features [91] as well as

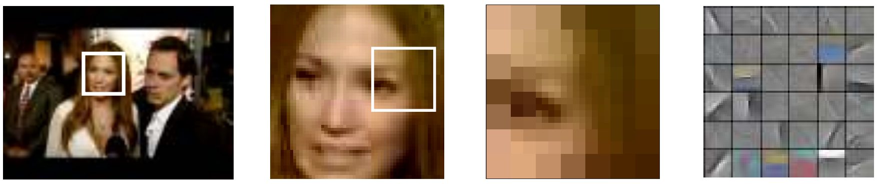


**Fig. 3.2. Left:** A typical video frame with the estimated target region highlighted. To cope with the large image size our system considers only the target region at each time step. **Centre left:** A close-up of the template extracted from the first frame. The template is compared to the target region by selecting a fixation point for comparison as shown. **Centre right:** A visualization of a single fixation. In addition to covering only a very small portion of the original frame, the image is foveated with high resolution near the centre and low resolution on the periphery to further reduce the dimensionality. **Right:** The most active filters of the first layer (factored)-RBM when observing the displayed location. The control pathway compares these features to the features active at the corresponding scene location in order to update the belief state.

features that are learned from data [276]. Core to these methods is the idea that saliency is determined by some type of novelty measure.

Our approach is different in that rather than identifying locally or globally novel features, our process identifies features which are useful for the task at hand. In our system the saliency signal for a location comes from a top down process which evaluates how well the features at that location enable the system to localize the target object. The work of [91] considers a similar approach to saliency by defining saliency to be the mutual information between the features at a location and the class label of an object being sought. However, in order to make their model tractable the authors are forced to use specifically engineered features and the approach is tightly coupled to their chosen task. Our system is able to handle arbitrary feature types, and although we consider only localization, our model is sufficiently general to be applied to identifying salient features for other goals. From the computer vision perspective, a top-down approach has been proposed in [138], a branch-and-bound method (faster that standard sliding window) to select the salient parts of the image for object detection and localization purposes.

Recently, a dynamic RBM state-space model was proposed in [239]. Both the implementation and intention behind that proposal are different from the approach discussed here. To the best of our knowledge, our approach is the first successful attempt to combine dynamic state estimation from gazes with online policy learning for gaze adaptation, using deep network network models of appearance. Many other dual-pathway architectures have been proposed in computational neuroscience, including [181] and [197], but we believe ours has the advantage that it is very simple, *modular* (with each module easily replaceable), suitable for large datasets and easy to extend.

Another interpretation of this work is as a new model for jointly learning to control eye movements (in smooth pursuit) and to estimate some unknown state about the world, mainly a target's position in some image. [174] is another example of a model performing estimation and control, but in a visual search task in which the estimation beliefs are non-linear and the control policy is greedy. An improvement of [174] has been proposed later in [49] by defining the same problem as a Partially Observable Markov Decision Process (POMDP) and applying a policy gradient algorithm to perform long term planning, based on an infomax reward. In [74], the authors propose a slightly different formulation based on continuous state representations (as opposed to a discretized continuous state space), and applied it to the problem of learning hand-eye coordination. In [125], the model is applied to the task of estimating the class of some input image (as opposed to the position of some target) from multiple fixations. Estimation is based on a non-parametric classifier, while control is random and based on a saliency map, derived from a model of natural images. An extension of the natural input memory to a Bayesian framework to decide where to saccade to next has been proposed in [24]. It is based on a bottom-up mechanism to build a salience map that inhibits the previously selected fixation points in order to avoid to explore always the same locations. Then, the acquired fragments of image are fused for multi-class recognition using the Bayesian naive model assumption. A distinguishing feature of our work is that we applied our model to video, as opposed to still images.

In this chapter, we are going to explore each component of the model reported in Figure 3.1. First, the identity pathway is analyzed in Section 3.3: The RBM appearance model used both for tracking and recognition is presented, then, we discuss how accumulate fixations over time with a multi-fixation RBM in order to perform classification. Section 3.4 concerns the control pathway. We present the state-space model, that is slightly different from the one presented in the previous chapter, because we introduce the concept of actions, *i.e.*, the fixation chosen at each step from the control method. Then, the control algorithm is discussed in Section 3.5. In Section 3.6, the full algorithm to perform inference summarizes the parts of the model discussed in the previous sections. Section 3.7 reports the experimental results for different application. Due to the generality of the approach, we show that the algorithm can be applied also in other context, not just for surveillance. Eventually, Section 3.8 highlights the potentiality of this work and the future work.

## 3.3 Identity Pathway

The identity pathway in our model mirrors the ventral pathway in neuroscience models. It is responsible for modeling the appearance of the target object and also, at a higher level, for classification. More specifically, we opt for a three layer architecture, followed by a classification module (see Figure 3.4). The first hidden layer aims at modeling the statistics of individual gaze instances or fixations, while the second hidden layer is trained to combine information about the relative position of many fixations with the first layer activations generated by those fixations, into a coherent representation. Finally, a classifier predicts the category of the tracked object based on the representation computed at the second hidden layer. Each step in this pathway is pre-trained greedily. It is worth noting that existing particle filtering and stochastic optimization algorithms could be used to train the RBMs online. We leave it as a possible extension of this work.

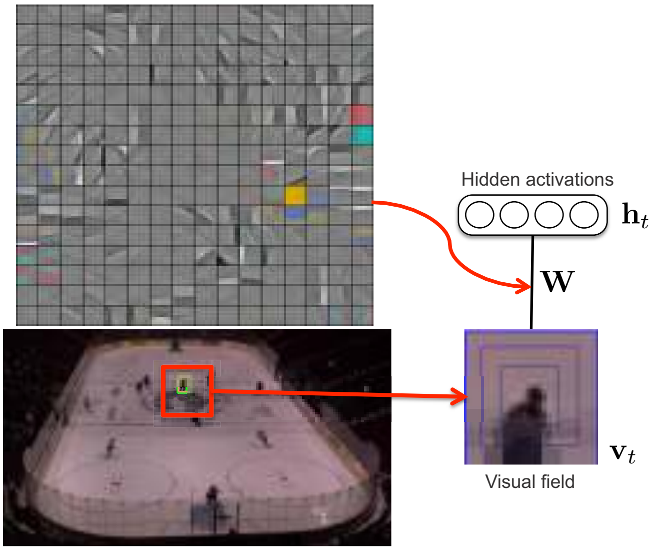


**Fig. 3.3.** An RBM senses a small foveated image derived from the video. The level of activation of each filter is recorded in the $\mathbf{h}_t$ units. The RBM weights (filters) $\mathbf{W}$ are visualized in the upper left. We currently pre-train these weights.

### 3.3.1 Appearance Model

The first hidden layer varies depending on the type of visual stimuli being modeled. For (approximately) binary inputs, we use a regular Restricted Boltzmann Machine (RBM) [87, 230]. Noting $\mathbf{v}_t$ as the observed fixation and $\mathbf{h}_t$ the RBM's binary hidden layer at time $t$, the assigned energy by the RBM is defined as

$$E(\mathbf{v}_t, \mathbf{h}_t) = -\mathbf{d}^\top \mathbf{h}_t - \mathbf{b}^\top \mathbf{v}_t - \mathbf{h}_t^\top \mathbf{W}\mathbf{v}_t$$

and probabilities are assigned through the Boltzmann distribution

$$p(\mathbf{v}_t, \mathbf{h}_t) = \frac{e^{-E(\mathbf{h}_t,\mathbf{v}_t)}}{Z} \ .$$

where $Z = \sum_{\mathbf{h}_t,\mathbf{v}_t} e^{-E(\mathbf{h}_t,\mathbf{v}_t)}$ is the partition function, usually intractable. Given a collection of randomly sampled fixations, the first layer RBM weights $\mathbf{W}$ and biases $\mathbf{d}, \mathbf{b}$ can be trained using contrastive divergence [106]. We refer the reader to [107] for a description of good practices in training RBMs. The end result is a hidden representation of the appearance of individual fixations $\mathbf{v}_t$[1], as

$$\mathbf{h}(\mathbf{v}_t) = [p(h_i = 1|\mathbf{v}_t)]_{i=1}^{H} = [\text{sigm}(d_i + \mathbf{W}_{i,:}\mathbf{v}_t)]_{i=1}^{H}$$

[1] More specifically, the first layer representation actually depends on the estimated track and the fixation point determined by gaze control (and implicitly, on the whole visual field), which yields the observed fixation $\mathbf{v}_t$. For simplicity of presentation, and to make our notation more compatible with the RBM literature, we ignore this dependency in the notation for now.

where the notation $\mathbf{W}_{i,:}$ refers to the $i^{\text{th}}$ row of the matrix $\mathbf{W}$ and $H$ is the number of hidden units of the first layer.

For stimuli better represented with real-valued inputs, such as color images, we used the factored RBM of [202] which is based on a different energy function that can be broken down in two parts:

$$E^c(\mathbf{v}_t, \mathbf{h}_t^c) = -(\mathbf{d}^c)^\top \mathbf{h}_t^c - \sum_{f=1}^{F} (\mathbf{P}_{f,:}\mathbf{h}_t^c)(\mathbf{C}_{f,:}\mathbf{v}_t)^2$$
$$E^m(\mathbf{v}_t, \mathbf{h}_t^m) = -(\mathbf{d}^m)^\top \mathbf{h}_t^m - (\mathbf{h}_t^m)^\top \mathbf{W}\mathbf{v}_t$$

where $F$ is the number of linear factors used to model the 3-way interactions between each hidden unit and pair of input units, $\mathbf{P}_{f,:}$ and $\mathbf{C}_{f,:}$ are the parameters associate to the $f$-th factor . A factored RBM models the data with two set of hidden units corresponding to the two energy functions defined above: $\mathbf{h}^m$ models the mean intensity of the each pixels independently, and $\mathbf{h}^c$ captures the pair-wise interactions between pixel values. Hybrid Monte Carlo (HMC) can then be used within a similar contrastive divergence procedure to train the parameters of this RBM [202].

The first hidden layer representation is defined using both sets of hidden units $\mathbf{h}(\mathbf{v}_t) = [\mathbf{h}^c(\mathbf{v}_t); \mathbf{h}^m(\mathbf{v}_t)]$, where

$$\mathbf{h}^c(\mathbf{v}_t) = [p(h_i^c = 1|\mathbf{v}_t)]_{i=1}^{H^c} = \left[\text{sigm}\left(d_i^c + \sum_{f=1}^{F} P_{f,i}(\mathbf{C}_{f,:}\mathbf{v}_t)^2\right)\right]_{i=1}^{H^c}$$
$$\mathbf{h}^m(\mathbf{v}_t) = [p(h_i^m = 1|\mathbf{v}_t)]_{i=1}^{H^m} = [\text{sigm}\,(d_i^m + \mathbf{W}_{i,:}\mathbf{v}_t)]_{i=1}^{H^m} \ .$$

This hidden layer can be understood as playing a similar role as the primary visual cortex (V1). In fact, when trained on patches of natural images, the factored RBM learns to extract a representation similar to the Gabor transform V1 neurons seem to be computing [202]. Also, much like the how in many neuroscience models V1 appears in both the ventral and dorsal pathways, the first hidden layer of our identity pathway appears in both the identity and control pathways (see Section 3.4.1).

### 3.3.2 Classification Model

Subsequent steps of the identity pathway are aimed at performing object recognition and classifying a sequence of fixations selected by the fixation policy.

To achieve this, we first implemented a multi-fixation RBM very similar to the one proposed in [143], where the binary variables $\mathbf{z}_t$ (see Figure 3.4) are introduced to encode the relative gaze location $\mathbf{a}_t$ within the multi-fixation RBM (a “1 in $K$” or “one hot” encoding of the gaze location was used for $\mathbf{z}_t$). This model sits on top of the appearance model described in the previous section.

The multi-fixation RBM uses the relative gaze location information in order to aggregate the first hidden layer representations $\mathbf{h}_t$ at $\Delta$ consecutive time steps into a single, higher level representation $\mathbf{h}_t^{[2]}$. More specifically, the energy function of the multi-fixation RBM is given by

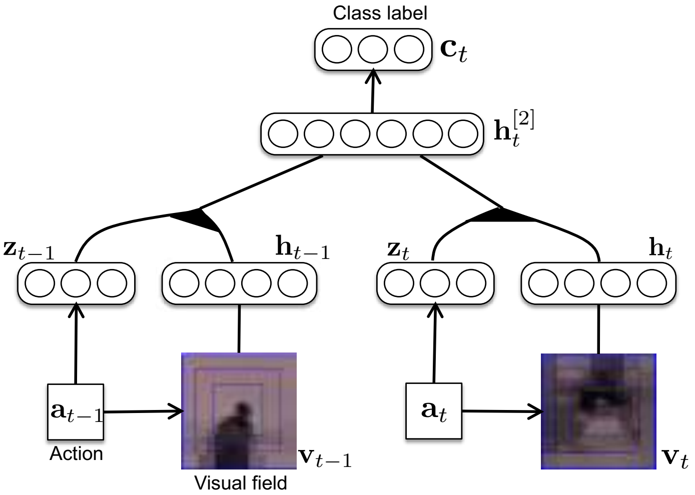


**Fig. 3.4.** Gaze accumulation and classification in the identity pathway. A multi-fixation RBM models the conditional distribution (given the gaze positions $\mathbf{a}_t$) of $\Delta$ consecutive hidden features $\mathbf{h}_t$, extracted by the first layer RBM. In this illustration, $\Delta = 2$. The multi-fixation RBM encodes the gaze position $\mathbf{a}_t$ in a "one hot" representation noted $\mathbf{z}_t$. The activation probabilities of the second layer hidden units $\mathbf{h}_t^{[2]}$ are used by a logistic regression classifier to predict the object's class.

$$E(\mathbf{h}_{t-\Delta+1:t}, \mathbf{z}_{t-\Delta+1:t}, \mathbf{h}_t^{[2]})$$
$$= -\mathbf{d}^{[2]^\top}\mathbf{h}_t^{[2]} - \sum_{i=1}^{\Delta}\left(\mathbf{b}^{[2]^\top}\mathbf{h}_{t-\Delta+i} + \sum_{f=1}^{F}(\mathbf{P}_{f,:}^{[2]}\mathbf{h}_t^{[2]})(\mathbf{W}_{f,:}^{[2]}\mathbf{h}_{t-\Delta+i})(\mathbf{V}_{f,:}^{[2]}\mathbf{z}_{t-\Delta+i})\right) .$$

where the learning parameters are the biases $dv^{[2]}$ and $\mathbf{b}^{[2]}$ and the weights $\mathbf{W}^{[2]}$, $\mathbf{P}^{[2]}$ and $\mathbf{V}^{[2]}$. From this energy function, we define a distribution over $\mathbf{h}_{t-\Delta+1:t}$ and $\mathbf{h}_t^{[2]}$ (conditioned on $\mathbf{z}_{t-\Delta+1:t}$) through the Boltzmann distribution

$$p(\mathbf{h}_{t-\Delta+1:t}, \mathbf{h}_t^{[2]}|\mathbf{z}_{t-\Delta+1:t}) = \frac{e^{-E(\mathbf{h}_{t-\Delta+1:t}, \mathbf{z}_{t-\Delta+1:t}, \mathbf{h}_t^{[2]})}}{Z(\mathbf{z}_{t-\Delta+1:t})} \quad , \tag{3.1}$$

where the normalization constant $Z(\mathbf{z}_{t-\Delta+1:t})$ ensures that Equation 3.1 sums to 1. To sample from this distribution, one can use Gibbs sampling by alternating between sampling the top-most hidden layer $\mathbf{h}_t^{[2]}$ given all individual processed gazes $\mathbf{h}_{t-\Delta+1:t}$ and vice versa. To train the multi-fixation RBM, we collect a training set consisting of sequences of $\Delta$ pairs $(\mathbf{h}_t, \mathbf{z}_t)$, obtained by randomly selecting $\Delta$ fixation points and computing the associated $\mathbf{h}_t$. These sets are extracted from a collection of images in which the object to detect has been centered. Unsupervised learning using contrastive divergence can then be performed on this training set. See [143] for more details.

The main difference between this multi-fixation RBM and the one described in [143] is that $\mathbf{h}_t^{[2]}$ does not explicitly model the class label $\mathbf{c}_t$. Instead, a multinomial logistic regression classifier is trained separately to predict $\mathbf{c}_t$ from the aggregated representation in $\mathbf{h}_t^{[2]}$. In this way, the multi-fixation RBM can be trained on unlabeled data and thus independently from the recognition task. Specifically, we use the vector of activation probabilities of the hidden units $h_{t,j}^{[2]}$ in $\mathbf{h}_t^{[2]}$, conditioned on $\mathbf{h}_{t-\Delta+1:t}$ and $\mathbf{z}_{t-\Delta+1:t}$, as the aggregated representation,

$$\begin{aligned} p(h_{t,j}^{[2]} &= 1|\mathbf{h}_{t-\Delta+1:t}, \mathbf{z}_{t-\Delta+1:t}) \\ &= \operatorname{sigm}\left( d_j + \sum_{i=1}^{\Delta}\sum_{f=1}^{F} \mathbf{P}_{f,j}^{[2]} (\mathbf{W}_{f,:}^{[2]}\mathbf{h}_{t-\Delta+i})(\mathbf{V}_{f,:}^{[2]}\mathbf{z}_{t-\Delta+i}) \right) \quad . \end{aligned}$$

To improve the estimate the class variable $\mathbf{c}_t$ over time, we accumulate the classification decisions at each time step. In particular, the class decision at each time step maximizes over $c$ the probability $p(c|\mathbf{c}_{0:t}) = \frac{\sum_{t'=0}^{t} I(\mathbf{c}_{t'}=c)}{t}$, where $I(\cdot)$ is the indicator function. We experimented with predicting the class label independently at each time step, but found the multi-fixation module to increase classification accuracy.

Note that the process of pursuit (tracking) is essential to classification. As the target is tracked, the algorithm fixates at locations near the target's estimated location. The size and orientation of these fixations also depends on the corresponding state estimates. The tracking estimates provide the locations where the algorithm gathers the gazes for classification. $\Delta$ gaze positions are randomly selected given the tracking estimates (one for each time step). This random selection is very important when the tracking policy has converged to a specific gaze. In that case, the selected gazes are similar, thus the multi-fixation RBM representation will converge to a single-fixation RBM, decreasing the classification accuracy. It should also be pointed out that instead of using random fixations, one could again use the control strategy proposed in this paper to decide where to look with respect to the track estimate so as to reduce classification uncertainty. We leave the implementation of this extra attentional mechanism for future work.

## 3.4 Control Pathway

The control pathway mirrors the responsibility of the dorsal pathway in human visual processing. It tracks the state of the target (position, speed, etc.) and normalizes the input so that other modules need not account for these variations. At a higher level, it is responsible for learning an attentional strategy which maximizes the amount of information learned with each fixation. The structure of the control pathway is shown in Figure 3.5, where it is easy to identify the standard state space evolution that includes the variables $\mathbf{x}$ and $\mathbf{h}$ and the control part including the variables $\mathbf{b}$, $\mathbf{r}$ and $\mathbf{a}$.

### 3.4.1 State-space model

As described in the previous chapter, the standard approach to image tracking is based on the formulation of Markovian, nonlinear, non-Gaussian state-space

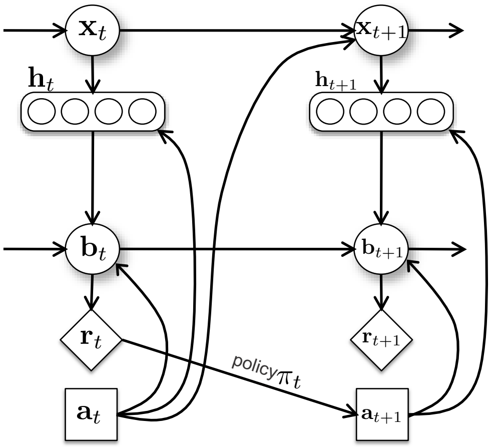


**Fig. 3.5.** Influence diagram for the control pathway. The true state of the tacked object $\mathbf{x}_{t+1}$, generates some set of features $\mathbf{h}_{t+1}$, in the identity pathway. These features depend on the action chosen at time $t+1$ and are used to update the belief state $\mathbf{b}_{t+1}$. Statistics of the belief state are collected to compute the reward $r_{t+1}$, which is used to update the policy for the next time step.

models, which are solved with approximate Bayesian filtering techniques. Note that in this chapter we use $\mathbf{x}_t$ instead of $\mathbf{X}_t$, because it simplifies the notation when tracking a single object $\mathbf{x}_t^1$. The proposed model borns to perform tracking of a single object at each time, because we wish to be consistent with the human perception system. In fact, recent cognitive models claim that humans can focus his/her attention only on one object at each time.

This signal has initial distribution $p\left(\mathbf{x}_0\right)$ and transition equation $p\left(\mathbf{x}_{t+1}\middle|\,\mathbf{x}_t, \mathbf{a}_t\right)$. Here $\mathbf{a}_t \in \mathcal{A}$ denotes an action at time $t$, defined on a compact set $\mathcal{A}$. For descrete policies $\mathcal{A}$ is finite whereas for continuous policies $\mathcal{A}$ is a region in $\mathbb{R}^2$. The observations $\{\mathbf{h}_t \in \mathcal{H}; t \in \mathbb{N}\}$, are assumed to be conditionally independent given the process state $\{\mathbf{x}_t; t \in \mathbb{N}\}$. Note that from the state space model perspective the observations are the hidden units of the first hidden layer of the appearance model in the identity pathway. In summary, the state-space model is described by the following distributions:

$$\begin{array}{rl} p\left(\mathbf{x}_0\right) & \\ p\left(\mathbf{x}_{t+1}\middle|\,\mathbf{x}_t, \mathbf{a}_t\right) & \text{for } t \geq 1 \ , \\ p\left(\mathbf{h}_{t+1}\middle|\,\mathbf{x}_{t+1}, \mathbf{a}_{t+1}\right) & \text{for } t \geq 1 \ . \end{array}$$

For the transition model, we adopt a classical autoregressive process. For the observation model, we follow common practice in image tracking and define it in terms of the distance of the observations from a template $\tau$,

$$p\left(\mathbf{h}_{t+1}|\mathbf{x}_{t+1}, \mathbf{a}_{t+1}\right) \propto e^{-d(\mathbf{h}(\mathbf{x}_{t+1}, \mathbf{a}_{t+1}), \tau)} \ ,$$

where $d(\cdot, \cdot)$ denotes a distance metric and $\tau$ an object template (for example, a color histogram or spline). Notice how we have now changed the notation from

$\mathbf{h}(\mathbf{v}_{t+1})$ to $\mathbf{h}(\mathbf{x}_{t+1}, \mathbf{a}_{t+1})$, to emphasize that the hidden unit activations are actually driven by the attentional policy, which in turn generates the fixation $\mathbf{v}_{t+1}$ in the first layer RBM.

In this model, the observation $\mathbf{h}(\mathbf{x}_{t+1}, \mathbf{a}_{t+1})$ is a function of the current state hypothesis and the selected action. The difficulty with this approach is eliciting a good template. Often color histograms or splines are insufficient. For this reason, we construct a template as follows. First, optical flow is used to detect new object candidates entering the visual scene. Second, we extract a region around the target to use as a visual template, as shown in Figure 3.2. The same figure also shows a typical foveated observation (higher resolution in the centre and lower in the periphery of the gaze) and the receptive fields for this observation learned beforehand with an RBM. The control algorithm is used to learn which parts of the template are most informative, either by picking from among a predefined set of fixation points, or by using a continuous policy. Finally, we define the likelihood of each observation directly in terms of the distance of the hidden units of the RBM $\mathbf{h}(\mathbf{x}_{t+1}, \mathbf{a}_{t+1})$, to the hidden units of the corresponding template region $\mathbf{h}(\mathbf{x}_1, \mathbf{a}_1 = k)$. That is,

$$p\left(\mathbf{h}_{t+1} | \mathbf{x}_{t+1}, \mathbf{a}_{t+1} = k\right) \propto e^{-d(\mathbf{h}(\mathbf{x}_{t+1}, \mathbf{a}_{t+1}=k), \mathbf{h}(\mathbf{x}_1, \mathbf{a}_1=k))} \quad .$$

The above template is static, but conceivably one could adapt it over time.

Our aim is to estimate recursively in time the *posterior distribution* $p(\mathbf{x}_{0:t+1} | \mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1})$ and its associated features, including the marginal distribution $\mathbf{b}_{t+1} \triangleq p\left(\mathbf{x}_{t+1} | \, \mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right)$ – known as the *filtering distribution* or *belief state.* This distribution satisfies the following recurrence:

$$\mathbf{b}_{t+1} \propto p(\mathbf{h}_{t+1} | \mathbf{x}_{t+1}, \mathbf{a}_{t+1}) \int p(\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{a}_t) p(d\mathbf{x}_t | \mathbf{h}_{1:t}, \mathbf{a}_{1:t}) \quad .$$

Except for standard distributions (*e.g.* Gaussian or discrete), this recurrence is intractable. We adopt particle filter to approximate the posterior distribution (see Section 3.6). Note that the recursive formulation is similar to the one discussed in the previous chapter, here we added the action variable.

### 3.4.2 Reward Function

A gaze control strategy specifies a policy $\pi(\cdot)$ for selecting fixation points. The purpose of this strategy is to select fixation points which maximize an instantaneous reward function $r_t(\cdot)$. The reward can be any desired behavior for the system, such as minimizing posterior uncertainty or achieving a more abstract goal. We focus on gathering observations so as to minimize the uncertainty in the estimate of the filtering distribution, $r_{t+1}(\mathbf{a}_{t+1} | \mathbf{b}_{t+1}) \triangleq u[\widetilde{p}(\mathbf{x}_{t+1} | \mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1})]$. More specifically, this reward is a function of the variance of the importance weights $w_{t+1}$ of the particle filter approximation of the belief state, that is, $r_{t+1}(\mathbf{a}_{t+1} | \mathbf{b}_{t+1}) = \sum_{i=1}^{N} (w_{t+1}^{(i)})^2$ (see Section 3.6 for more details).

It is also useful to consider the cumulative reward

$$R_T = \sum_{t=1}^{T} r_t(\mathbf{a}_t | \mathbf{b}_t) \quad ,$$

which is the sum of the instantaneous rewards which have been received up to time $T$. The gaze control strategies we consider are all "no-regret" which means that the average gap between our cumulative reward and the cumulative reward from always picking the single best action goes to zero as $T \to \infty$.

In our model each action is a different gaze location and the objective is to choose where to look so as to minimize the uncertainty about the belief state.

## 3.5 Gaze control

We compare two different strategies for learning the gaze selection policy. Here we proposed to use a model that learns the gaze selection policy with a portfolio allocation algorithm called Hedge [14, 88]. Hedge requires knowledge of the rewards for all actions at each time step, which is not realistic when gazes must be preformed sequentially, since the target object will move between fixations. For this reason, it is also categorized as *full information game* However, several improvements can be done using *partial information games*. For example, EXP3 and its extensions [13] require knowledge of the reward only for the action selected at each time step. They have been sucessufully employed in the future work of this thesis in [67]. Hedge and EXP3 learn gaze selection policies which choose among a discrete set of predetermined fixation points. Another extension can be to learn a *continuous policy* by estimating the reward surface using a Gaussian Process [204] with Bayesian optimization [45].

In this work, we compare Hedge to two baseline methods. The following sections describe each of these approaches in more detail.

**Baselines.** We consider two baseline strategies, which we call random and circular. The random strategy samples fixation points uniformly at random from a small discrete set of possibilities. The circular strategy also uses a small discrete set of fixation points and cycles through them in a fixed order.

**Hedge.** To use Hedge [14, 88] for gaze selection we must first discretize the action space by selecting a fixed finite number of possible fixation points. Hedge maintains an importance weight $G(i)$ for each possible fixation point and uses them to form a stochastic policy at each time step. An action is selected according to this policy and the reward for each possible action is observed. These rewards are then used to update the importance weights and the process repeats. Pseudo code for Hedge is shown in Algorithm 2.

## 3.6 Algorithm

Since the belief state cannot be computed analytically, we will adopt particle filtering to approximate it. The full algorithm is shown in Algorithm 3.

We refer readers to [69] for a more in depth treatment of these sequential Monte Carlo methods. Assume that at time $t$ we have $N \gg 1$ particles (samples) $\{\mathbf{x}_{0:t}^{(i)}\}_{i=1}^{N}$ distributed according to $p\left(d\mathbf{x}_{0:t}|\mathbf{h}_{1:t}, \mathbf{a}_{1:t}\right)$. We can approximate this belief state with the following empirical distribution

**Algorithm 2**: Hedge

```
Require: $\gamma > 0$
Require: $G_0(i) \leftarrow 0$   foreach $i \in \mathcal{A}$
  for $t = 1, 2, \ldots$ do
    for $i \in \mathcal{A}$ do
      $p_t(i) \leftarrow \frac{\exp \gamma G_{t-1}(i)}{\sum_{j \in \mathcal{A}} \exp \gamma G_{t-1}(j)}$
    $\mathbf{a}_t \sim (p_t(1), \ldots, p_t(|\mathcal{A}|))$ {sample an action from the distribution $(p_t(k))$}
    for $i \in \mathcal{A}$ do
      $r_t(i) \leftarrow r_t(i|\mathbf{b}_t)$
      $G_t(i) \leftarrow G_{t-1}(i) + r_t(i)$
```

$$\widehat{p}\left(d\mathbf{x}_{0:t}|\mathbf{h}_{1:t}, \mathbf{a}_{1:t}\right) \triangleq \frac{1}{N} \sum_{i=1}^{N} \delta_{\mathbf{x}_{0:t}^{(i)}}\left(d\mathbf{x}_{0:t}\right) \quad .$$

Particle filters combine sequential importance sampling with a selection scheme designed to obtain $N$ new particles $\{\mathbf{x}_{0:t+1}^{(i)}\}_{i=1}^{N}$ distributed approximately according to $p\left(d\mathbf{x}_{0:t+1}|\mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right)$.

### 3.6.1 Importance sampling step

This section reports a formalization similar to Section 2.2, but now we have an additional variable, that is, the action $\mathbf{a}_t$. For the sake of clarity, we report here the formulation of importance sampling similar to Section 2.2.

The joint distributions $p\left(d\mathbf{x}_{0:t-1}|\mathbf{h}_{1:t}, \mathbf{a}_{1:t}\right)$ and $p\left(d\mathbf{x}_{0:t+1}|\mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right)$ are of different dimension. We first modify and extend the current paths $\mathbf{x}_{0:t}^{(i)}$ to obtain new paths $\widetilde{\mathbf{x}}_{0:t+1}^{(i)}$ using a proposal kernel $q_{t+1}\left(d\widetilde{\mathbf{x}}_{0:t+1}|\mathbf{x}_{0:t}, \mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right)$. As our goal is to design a sequential procedure, we set

$$q_{t+1}\left(d\widetilde{\mathbf{x}}_{0:t+1}|\,\mathbf{x}_{0:t}, \mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right) = \delta_{\mathbf{x}_{0:t}}\left(d\widetilde{\mathbf{x}}_{0:t}\right) q_{t+1}\left(d\widetilde{\mathbf{x}}_{t+1}|\,\widetilde{\mathbf{x}}_{0:t}, \mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right) ,$$

that is $\widetilde{\mathbf{x}}_{0:t+1} = (\mathbf{x}_{0:t}, \widetilde{\mathbf{x}}_{t+1})$. The aim of this kernel is to obtain new paths whose distribution

$$q_{t+1}\left(d\widetilde{\mathbf{x}}_{0:t+1}|\mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right) = p\left(d\widetilde{\mathbf{x}}_{0:t}|\mathbf{h}_{1:t}, \mathbf{a}_{1:t}\right) q_{t+1}\left(d\widetilde{\mathbf{x}}_{t+1}|\,\widetilde{\mathbf{x}}_{0:t}, \mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right) ,$$

is as "close" as possible to $p\left(d\widetilde{\mathbf{x}}_{0:t+1}|\mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right)$. Since we cannot choose $q_{t+1}\left(d\widetilde{\mathbf{x}}_{0:t+1}|\mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right) = p\left(d\widetilde{\mathbf{x}}_{0:t+1}|\mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right)$ because this is the quantity we are trying to approximate in the first place, it is necessary to weight the new particles so as to obtain consistent estimates. We perform this "correction" with importance sampling, using the weights

$$\widetilde{w}_{t+1} = \widetilde{w}_t \frac{p\left(\mathbf{h}_{t+1}|\widetilde{\mathbf{x}}_{t+1}, \mathbf{a}_{t+1}\right) p\left(d\widetilde{\mathbf{x}}_{t+1}|\widetilde{\mathbf{x}}_{0:t}, \mathbf{a}_t\right)}{q_{t+1}\left(d\widetilde{\mathbf{x}}_{t+1}|\,\widetilde{\mathbf{x}}_{0:t}, \mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right)} \quad .$$

The choice of the transition prior as proposal distribution is by far the most common one. In this case, the importance weights reduce to the expression for the likelihood. However, it is possible to construct better proposal distributions,

**Algorithm 3**: Particle filtering algorithm with gaze control for full information policies. Note that for partial information policies the algorithm does not have to iterate on all the $K$ gazes.

```
1. Initialization
  for $i = 1$ to $N$ do
    $\mathbf{x}_0^{(i)} \sim p(\mathbf{x}_0)$
  Initialize the policy $\pi_1(\cdot)$ {How this is done depends on the control strategy}
for $t = 0 \ldots$ do
  2. Importance sampling
    for $i = 1$ to $N$ do {Predict the next state}
      $\widetilde{\mathbf{x}}_{t+1}^{(i)} \sim q_{t+1}\left(d\mathbf{x}_{t+1}^{(i)} \middle| \widetilde{\mathbf{x}}_{0:t}^{(i)}, \mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right)$
      $\widetilde{\mathbf{x}}_{0:t+1}^{(i)} \leftarrow \left(\mathbf{x}_{0:t}^{(i)}, \widetilde{\mathbf{x}}_{t+1}^{(i)}\right)$
    for $k = 1$ to $K$ do
      for $i = 1$ to $N$ do {Evaluate the importance weights for each gaze position}
        $\widetilde{w}_{t+1}^{(i)}(k) \leftarrow \frac{p\left(\mathbf{h}_{t+1}|\widetilde{\mathbf{x}}_{t+1}^{(i)}, \mathbf{a}_{t+1} = k\right) p\left(\widetilde{\mathbf{x}}_{t+1}^{(i)}|\widetilde{\mathbf{x}}_{0:t}^{(i)}, \mathbf{a}_t\right)}{q_{t+1}\left(\widetilde{\mathbf{x}}_{t+1}^{(i)} \middle| \widetilde{\mathbf{x}}_{0:t}^{(i)}, \mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right)}$
      for $i = 1$ to $N$ do {Normalize the importance weights}
        $w_{t+1}^{(i)}(k) \leftarrow \frac{\widetilde{w}_{t+1}^{(i)}(k)}{\sum_{j=1}^N \widetilde{w}_{t+1}^{(j)}(k)}$
  3. Gaze control
    for $k = 1$ to $K$ do {Compute the reward for each gaze position}
      $r_{t+1}(k) = \sum_{i=1}^N (w_{t+1}^{(i)}(k))^2$ {Receive reward for the chosen action}
    Incorporate $r_{t+1}$ into the policy to create $\pi_{t+2}(\cdot)$
    Select an action $k^\star \sim \pi_{t+1}(\cdot)$
  4. Selection
    Resample with replacement $N$ particles $\left(\mathbf{x}_{0:t+1}^{(i)}; i = 1, \ldots, N\right)$ from the set
    $\left(\widetilde{\mathbf{x}}_{0:t+1}^{(i)}; i = 1, \ldots, N\right)$ according to the normalized weights $w_{t+1}^{(i)}(k^\star)$
```

which make use of more recent observations, using object detectors [178], saliency maps [114], optical flow, and approximate filtering methods such as the unscented particle filter. One could also easily incorporate strategies to manage data association and other tracking related issues. After normalizing the weights, $w_{t+1}^{(i)} = \frac{\widetilde{w}_{t+1}^{(i)}}{\sum_{j=1}^N \widetilde{w}_{t+1}^{(j)}}$, we obtain the following estimate of the filtering distribution:

$$\widetilde{p}\left(d\mathbf{x}_{0:t+1}|\mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1}\right) = \sum_{i=1}^N w_{t+1}^{(i)} \delta_{\widetilde{\mathbf{x}}_{0:t+1}^{(i)}}\left(d\mathbf{x}_{0:t+1}\right) \quad .$$

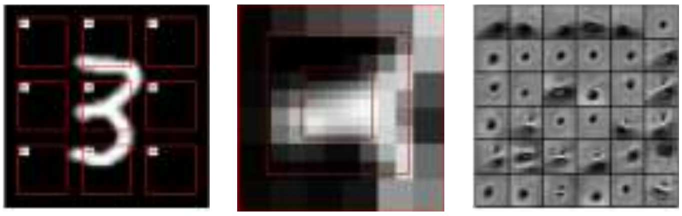


**Fig. 3.6. Left:** An example of a digit template used in the first experiment. The red boxes show the positions of each possible fixation point. **Centre:** The foveated fixation that corresponds to the fixation point G5. **Right:** The most active RBM filters when fixating on G5.

Finally a selection step is used to obtain an "unweighted" approximate empirical distribution $\hat{p}(d\mathbf{x}_{0:t+1}|\mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1})$ of the weighted measure $\tilde{p}(d\mathbf{x}_{0:t+1}|\mathbf{h}_{1:t+1}, \mathbf{a}_{1:t+1})$. The basic idea is to discard samples with small weights and multiply those with large weights. The use of a selection step is key to making the sequential Monte Carlo procedure effective; see [69] for details on how to implement this black box routine.

## 3.7 Experiments

In this section we report the results of running our system on several different videos sequences including both synthetic and real world data. We consider the full information scenario and demonstrate that a learned attentional policy outperforms the baseline strategies both in terms of tracking performance as well as classification accuracy.

Three experiments are carried out to evaluate quantitatively and qualitatively the proposed approach. The first experiment provides comparisons between Hedge and the baseline policies. The second experiment, on a similar synthetic dataset, demonstrates how the approach can handle large variations in scale, occlusion and multiple targets. The final experiment is a demonstration of tracking and classification performance on several real videos. For the synthetic digit videos, we trained the first-layer RBMs on the foveated images, while for the real videos we trained factored-RBMs on foveated natural image patches [202].

The first experiment uses 10 video sequences (one for each digit) built from the MNIST dataset [145]. Each sequence contains a moving digit and static digits in the background (to create distractions). The gaze template had $K = 9$ gaze positions, chosen so that gaze G5 was at the centre as shown in Figure 3.6, and the objective is to track and recognize the moving digit (see Figure 3.7). The location of the template was initialized with optical flow.

We compare the leaned policy (Hedge) against two baselines: the random policy and the circular policy (see Section 3.5). The Bhattacharyya distance has been used in the specification of the observation model. A multi-fixation RBM was trained to map the first layer hidden units of three consecutive time steps into a second

| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **Avg** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Hedge** | **1.2**<br>(1.2) | **3.0**<br>(2.0) | **2.9**<br>(1.0) | **2.2**<br>(0.7) | **1.0**<br>(1.9) | **1.8**<br>(1.9) | **3.8**<br>(1.0) | **3.8**<br>(1.5) | **1.5**<br>(1.7) | **3.8**<br>(2.8) | **2.5**<br>(1.6) |
| **Circular** | 18.2<br>(29.6) | 536.9<br>(395.6) | 104.4<br>(69.7) | 2.9<br>(2.2) | 201.3<br>(113.4) | 4.6<br>(4.0) | 5.6<br>(3.1) | 64.4<br>(45.3) | 142.0<br>(198.8) | 144.6<br>(157.7) | 122.5<br>(101.9) |
| **Random** | 41.5<br>(54.0) | 410.7<br>(329.4) | 3.2<br>(2.0) | 3.3<br>(2.4) | 42.8<br>(60.9) | 6.5<br>(9.6) | 5.7<br>(3.2) | 80.7<br>(48.6) | 38.9<br>(50.6) | 225.2<br>(241.6) | 85.9<br>(80.2) |

**Table 3.1.** Tracking error (in pixels) on several video sequences using different policies for gaze selection.

| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **Avg** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Hedge** | 95.62% | **100.00%** | **99.66%** | 99.33% | **99.66%** | **100.00%** | **100.00%** | **98.32%** | **97.98%** | **89.56%** | **98.01%** |
| **Circular** | **99.33%** | **100.00%** | 98.99% | 94.95% | 5.39% | 98.32% | 0.00% | 29.63% | 52.19% | 0.00% | 57.88% |
| **Random** | 98.32% | **100.00%** | 96.30% | **99.66%** | 29.97% | 96.30% | 89.56% | 22.90% | 12.79% | 13.80% | 65.96% |

**Table 3.2.** Classification accuracy on several video sequences using different policies for gaze selection.

hidden layer, and we trained a logistic regressor to further map to the 10 digit classes. We used the transition prior as proposal for the particle filter.

Tables 3.1 and 3.2 report the comparison results. Tracking accuracy was measured in terms of the mean and standard deviation (in brackets) over time of the distance between the target ground truth and the estimate; measured in pixels. The analysis highlights that the error of the Hedge policy is always below the error of the other policies. In most of the experiments, the tracker fails when an occlusion occurs for the deterministic and the random policies, while the learned policy is successful. This is very clear in the videos at: `http://www.youtube.com/user/anonymousTrack`

The loss of track for the simple policies is mirrored by the high variance results in Table 3.1 (experiments 0, 1, 4, and so on). The average mean and standard deviations (last column of Table 3.1) make it clear that the proposed strategy for learning a gaze policy can be of enormous benefit. The improvements in tracking performance are mirrored by improvements in classification performance.

Figure 3.7 provides further anecdotal evidence for the policy learning algorithm. The top sequence shows the target and the particle filter estimate of its location over time. The middle sequence illustrates how the policy changes over time. In particular, it demonstrates that hedge can effectively learn where to look in order to improve tracking performance (we chose this simple example as in this case it is obvious that the centre of the eight (G5) is the most reliable gaze action). The classification results over time are shown in the third row.

The second experiment addresses a similar video sequence, but tracking multiple targets. We instantiate an independent model for each target, thus the state space is disjoint. The image scale of each target changes significantly over time, so the algorithm has to be invariant with respect to these scale transformations. In this case, we used a mixture proposal distribution consisting of motion detectors and the transition prior. We also tested a saliency proposal but found it to be less effective than the motion detectors for this dataset. Figure 3.8 (top) shows some of the video frames and tracks. The videos allow one to better appreciate the performance of the multi-target tracking algorithm in the presence of occlusions.

Tracking and classification results for the real videos are shown in Figure 3.8 and the accompanying videos. We analyzed three different scenario: *hockey*, *surveil-*

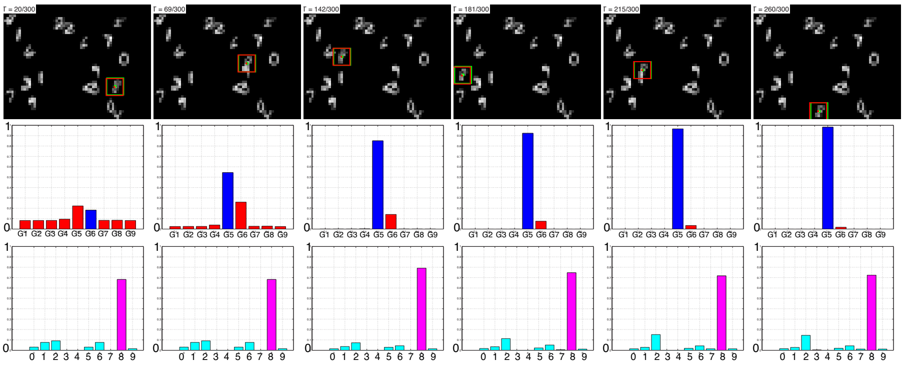


**Fig. 3.7.** Tracking and classification accuracy results with the learned policy. **First row:** position of the target and estimate over time. **Second row:** policy distribution over the 9 gazes; hedge clearly converges to the most reasonable policy. **Third row:** cumulative class distribution for recognition.

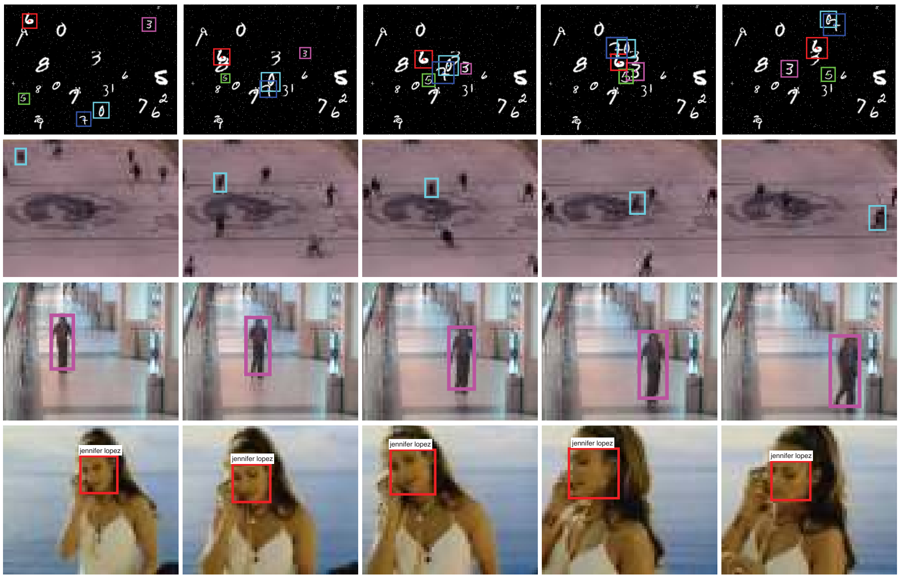


**Fig. 3.8. Top:** Multi-target tracking with occlusions and changes in scale on a synthetic video. **Middle and bottom:** Tracking in real video sequences.

*lance*, *face*. The *hockey* scenario consists on a video of hockey players taken from a static camera. For the *surveillance* scenario, we extracted a video from popular public dataset for people detection and tracking, CAVIAR [1] . For the *face* scenario, we use the Youtube celebrity dataset from [130]. This data set consists of several videos of celebrities taken from Youtube and is challenging for tracking algorithms as the videos exhibit a wide variety of illuminations, expressions and face orientations. For these datasets we used a proposal based on detections as discussed in Section 2.2.2. The qualitative results shown in Figure 3.8 provides additional evidence that the proposed model is able to perform tracking on real data. Have a look also at the available videos on our youtube channel.

## 3.8 Conclusions

We have proposed a decision-theoretic probabilistic graphical model for joint classification, tracking and planning. The experiments demonstrate the significant potential of this approach. We examined several strategies for gaze control in a full information setting, where all the available gazes are evaluated.

There are many routes for further exploration. In this work we pre-trained the appearance model. However, existing particle filtering and stochastic optimization algorithms could be used to train the RBMs online. Following the same methodology, we should also be able to adapt and improve the target templates and proposal distributions over time. This is essential to extend the results to long video sequences where the object undergoes significant transformations (*e.g.*, as is done in the predator tracking system [123]).

Deployment to more complex video sequences will require more careful and thoughtful design of the proposal distributions, transition distributions, control algorithms, template models, data-association and motion analysis modules. Fortunately, many of the solutions to these problems have already been engineered in the computer vision, tracking and online learning communities. Admittedly, much work remains to be done. See Section 2.1 of the previous chapter for the discussion about how to deal with these problems.

Saliency maps are ubiquitous in visual attention studies. Here, we simply used standard saliency tools and motion flow in the construction of the proposal distributions for particle filtering. There might be better ways to exploit the saliency maps, as neurophysiological experiments seem to suggest [99].

One of the most interesting avenues for future work is the construction of more abstract attentional strategies. In this work, we focused on attending to regions of the visual field, but clearly one could attend to subsets of receptive fields or objects in the deep appearance model.

A closer examination of the exploration/exploitation tradeoff in the tracking setting is in order. For instance, the methods we considered assume that future rewards are independent of past actions. This assumption is clearly not true in our setting, since choosing a long sequence of very poor fixation points can lead to tracking failure. We can potentially solve this problem by incorporating the current tracking confidence into the gaze selection strategy. This would allow the exploration/exploitation trade off to be explicitly modulated by the needs of the tracker, *e.g.*, after choosing a poor fixation point the selection policy could be adjusted temporarily to place extra emphasis on exploiting good fixation points until confidence in the target location has been recovered. Contextual bandits provide a framework for integrating and reasoning about this type of side-information in a principled manner.

# Part II

# Person Re-identification for Multi-camera Tracking

# 4

# Person Re-identification

In the previous part of this thesis, several tracking methods have been discussed, without taking particular care of the characteristics that a real environment can have, that is usually monitored by multiple cameras. This chapter is devoted to reply to the following questions: What happens in case of multiple cameras? How can a system deal with a camera network? In other words, when multiple independent tracking models are available (Fig. 4.1), we need also to link the results through cameras in order to keep persistent identifiers in the camera network (the variable $\mathbf{P}_{t+1}$ in Fig. 4.1). In literature, this issue is called *person re-identification*. It is worth noting that we can apply the same methods in order to perform *person re-acquisition*, *i.e.*, associating IDs in case of tracking failure in the same camera.

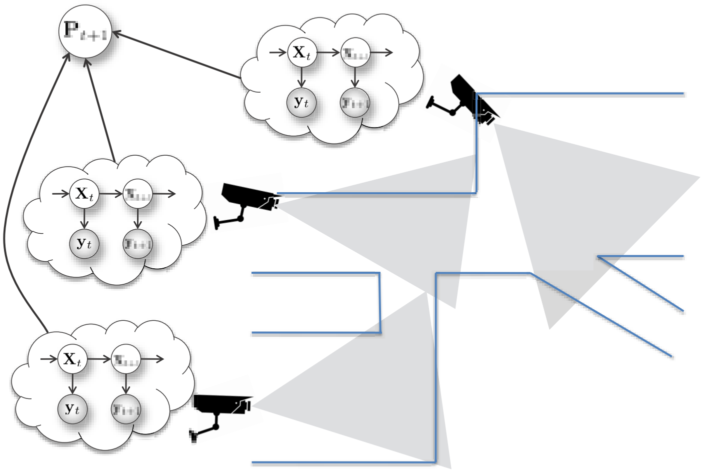


**Fig. 4.1.** Multi-camera model for person re-identification.

## 4.1 Introduction

When dealing with non-overlapped camera views, the person re-identification and re-acquisition methods are focused in modeling the human appearance. In general, characterizing the human appearance in surveillance is a hard task: most of the time people are captured by different low resolution cameras, under occlusions conditions, badly illuminated, and in different poses. The modeling problem becomes even harder when human descriptions serve as signatures in a recognition scenario. For example, in the re-identification problem, personal signatures must be matched across hundreds of candidates which have been captured in various locations and/or in different moments. The classical multi-target tracking issue is another case where individual descriptions are exploited to ensure consistent tracks across time. In this context, a robust modeling of the entire body appearance of a person is mandatory, especially when other classical biometric cues (face, gait) are not available or difficult to catch, due to the sensors' scarce resolution or low frame-rate.

In this chapter, we propose a pipeline that can be used as standard for the re-identification problem. We will see that the core of the pipeline is the descriptor used for characterizing the human appearance, and thus we propose three feature-based descriptors that reach the state-of-the-art results of the investigated problem. Such descriptors may be cast naturally in a re-identification context. In addition, we will see that one of them is also particularly suited as person model for tracking.

The first descriptor that actually also delineates the pipeline for re-identification is dubbed *Symmetry-Driven Accumulation of Local Features* (SDALF). The pipeline is composed by six steps: 1) images are gathered from a tracker, 2) then redundant information is discarded. 3) The person is segmented in foreground/background regions and 4) then in symmetric and asymmetric parts. 5) The descriptor is extracted and accumulated over time. 6) Finally, the matching between the probe signature and the ones stored in the gallery set (or database) is carried out. The novelties we present in this thesis concern mainly the last three steps. SDALF is a (a)symmetry-based description of the human body, and it was inspired by the fact that most natural objects and phenomena manifest symmetry in some form, so detecting and characterizing symmetry is a natural way to understand the structure of objects. The Gestalt psychology school [133] considers symmetry as a fundamental principle of perception: symmetrical elements are more likely integrated into one coherent object than asymmetric regions. This principle has been also largely exploited in computer vision for characterizing salient parts of a structured object [54, 147, 206, 209]. In SDALF, asymmetry principles allow to segregate meaningful body parts (head, upper body, lower body), whereas symmetry criteria help in extracting features from the actual human body, pruning out distracting background clutter. The idea is that features near the vertical axis of symmetry are weighted more than those that are far from it, ensuring to get information from the internal part of the body, trusting less the peripheral portions. This perceptual part localization is robust as it operates at dramatic low resolution (up to 11 × 22), under pose, viewpoint, and illumination changes.

Once parts have been localized, complementary aspects of the human body appearance are extracted in SDALF, highlighting: i) the global chromatic content, by the color histogram (see Fig. 4.6(c)); ii) the per-region color displacement, employing Maximally Stable Colour Regions (MSCR) [85] (see Fig. 4.6(d)); iii) the presence of *Recurrent Highly Structured Patches* (RHSP), estimated by a novel per-patch similarity analysis (see Fig. 4.6(e)).

An important aspect of the proposed method is that it exploits the presence of multiple instances (images) of the same person for reinforcing its characterization. This occurs in several surveillance scenarios: to quote a few, human operators may employ Pan-Tilt-Zoom (PTZ) cameras to grab as many images of a suspect as possible; in whatever tracking approach, consecutive shots of a tracked individual are available, that help in revising the object model against appearance changes. SDALF takes into account these situations, collecting features from all the available pictures of an individual, thus augmenting the robustness and the expressiveness of its description.

After the signature has been built, our method adopts a simple, on-line distance minimization strategy to match a probe signature with the ones in the gallery set. We distinguish between single-shot and multi-shot modalities, where the former considers the case when the signature is built using only one image for each individual, while the latter uses multiple images.

In this chapter, we will show that it is possible to use SDALF as person descriptor for tracking similarly to the multi-shot modality. The idea is to build a signature for each tracked target (the template). Then, the signature is matched to a gallery set made by the current hypothesis of the particle filter. The matching procedure will output a similarity score, that is embedded in the observation model of the filter. Thus instead of doing deterministic matching, *i.e.*, it selects only the high-score matching, the tracker performs a probabilistic matching, *i.e.*, it keeps the score of each hypothesis. The template is also naturally updated with SDALF descriptor: multiple images are gathered over time in the multi-shot modality. The use of SDALF for target tracking gives additional evidence of the robustness of the proposed descriptor for the characterization for the human appearance.

The second and third proposed descriptors are called *Histogram Plus Epitome* (HPE) and its extension, *Asymmetry-driven HPE* (AHPE), respectively. These two descriptors follow the same pipeline proposed for SDALF. The main distinctions are: 1) the local and global features used in the descriptor, 2) the matching phase that works only in the multi-shot modality, and 3) only AHPE does employ the (a)symmetry principles to segment parts of the individual, instead HPE takes into account the whole foreground figure.

(A)HPE descriptor incorporates global and local statistical descriptions of the human appearance, focusing on the global chromatic content via a mean color histogram, and the presence of recurrent local patterns through epitomic analysis proposed by [118]. The former captures the chromatic information of an individual's appearance, condensing it in a widely accepted descriptor for re-identification. The latter is supported by the paradigm of object recognition by local features, called epitome [118], that encodes the pixels' local spatial layout with a set of frequently visible patches. Another advantage of the epitome is that it naturally accumulates images in a multi-shot descriptor. We then exploit the asymmetry-based segmentation proposed by SDALF in order to apply our signature as human part descriptor, giving rise to AHPE.

We test the proposed methods on the most challenging public datasets: ViPER [101], iLIDS for re-identification [275], and ETHZ [222], setting in most of the cases state-of-the-art performances. These datasets embed different challenges for the re-identification problem: pose, viewpoint and lighting variations, and occlusions. As further analysis, we propose a novel dataset, that is CAVIAR4REID [26], extracted from the CAVIAR repository [1]. The main characteristic of CAVIAR4REID is that it is very close to a real scenario of re-identification, that is, multiple images for each individuals and multi-camera. This allows to clearly understand the benefit of having multiple instances per person in a re-identification challenge. Moreover, we test the limit of SDALF by subsampling these dataset up to dramatic resolutions ($11 \times 22$ pixels). Exploiting SDALF as an appearance model for the tracking, we consider the widely-known CAVIAR sequence dataset and a Bayesian multi-target tracker. We will show that SDALF outperforms the classical object descriptors considered in the literature.

The rest of the chapter is organized as follows. In Section 4.2, the state of the art of re-identification is described, highlighting the differences of the existing methods with respect to our strategy. Section 4.3 details the re-identification pipeline with particular focus on the SDALF descriptor. Section 4.4 describes how signatures matching is performed. Section 4.5 report the use of SDALF for tracking. In Section 4.6, HPE and AHPE are analyzed. Several comparative results are reported in Section 4.7, and, finally, conclusions and future perspectives are discussed in Section 4.8.

## 4.2 Related Work

Re-identification methods that rely only on visual information are addressed as *appearance-based* techniques. Other approaches assume less general operative conditions: *geometry-based* techniques exploit geometrical constrains in a scenario with overlapped camera views [195, 239]. However, in surveillance applications this scenario is very uncommon. *Temporal-based* methods deal with non-overlapped views adding a temporal reasoning on the spatial layout of the monitored environment, in order to prune the candidate set to be matched [116, 162, 201]. The assumption is that people usually enter in few locations of the image, spending a certain, fixed period (learned beforehand) in the blind spots. In this thesis, we manly focus on the appearance-based techniques, because of its generality and because we think that characterizing the human appearance is the most challenging of the problems in person re-identification. For this reason, the review in this section will be focused on the appearance-based methods.

Appearance-based methods can be divided into two groups (see Table 4.1): the *learning-based* methods and the *direct* methods. The former group is characterized by the use of a training dataset of *different individuals* where the features and/or the policy for combining them that ensures high re-identification accuracies are analyzed [19, 102, 156, 175, 199, 222, 226, 275]. The underlying assumption is that the knowledge extracted from the training set could generalize to unseen examples.

| | Single-shot | Multiple-shot |
|---|---|---|
| *Learning-based* | [156, 175, 199, 222]<br>[19, 102, 275] | [226] |
| *Direct Methods* | [20]<br>**SDALF** | [39, 95, 104, 259]<br>**SDALF, (A)HPE** |

**Table 4.1.** Taxonomy of the existing re-identification methods.

In [175], local and global features are accumulated over time for each subject, and fed into a multi-class SVM for recognition and pose estimation, employing different learning schemes. Viewpoint invariance is instead the main issue addressed by [102]: spatial and color information are here combined using an ensemble of discriminant localized features and classifiers selected by boosting. In [156], pairwise dissimilarity profiles between individuals are learned and adapted for a nearest neighbor classification. Similarly, in [222], a high-dimensional signature composed by texture, gradient and color information is projected into a low-dimensional discriminant latent space by Partial Least Squares (PLS) reduction. An "unconventional" approach is proposed by [275], where the description of a person is enriched by contextual visual knowledge coming from the surrounding people that form a group. The method implies that a group association between two or more people holds in different locations of a given environment, and exploits novel visual group descriptors, embedding visual words into concentric spatial structures. Re-identification is cast as a binary classification problem (one vs. all) by [19] using Haar-like features and a part-based MPEG7 dominant color descriptor. In [199], the re-identification problem is reformulated as a ranking problem and an informative subspace is learned where the potential true match is given highest ranking. Ensemble RankSVM is proposed as ranking method, reducing significantly the memory requirements.

It is worth noting that the learning-based approaches are strongly dependent on the cardinality and the kind of training set. Such approaches may suffer of generalization problems so that they have to be frequently re-trained/updated, when facing real scenarios (*e.g.*, an airport), while the gallery set changes quickly and consistently (*e.g.*, new individuals entering into the monitored area).

The other class of approaches, the direct methods, does not consider training datasets of multiple people and rather work on each person independently [20, 39, 95, 104, 259], usually focusing on designing novel features for capturing the most distinguishing aspects of an individual. In [39], the bounding box of a pedestrian is equally subdivided into ten horizontal stripes, and the median HSL value is extracted in order to manage $x$-axis pose variations. These values, accumulated over different frames, generate a multiple signature. A spatio-temporal local feature grouping and matching is proposed by [95], considering ten consecutive frames for each person, and estimating a region-based segmented image. The same authors present a more expressive model, building a decomposable triangulated graph that captures the spatial distribution of the local descriptions over time, so as to allow a more accurate matching. In [259], the method consists in segmenting a pedestrian image into regions, and registering their color spatial relationship into a co-occurrence matrix. This technique proved to work well when pedestrians are seen from small variations of the point of view. In [104], the person re-identification scheme is based on the matching of SURF [25] interest points collected in several images during short video sequences. Covariance features, originally employed for pedestrian detection, are extracted from coarsely located body parts and tailored for re-identification purposes [20].

Considering the features employed for re-identification, in addition to color information which is universally adopted, several other features of interest are textures [102,199,222], edges [222], Haar-like features [19], interest points [95], image patches [102], and segmented regions [259]. These features, when not collected densely, can be extracted from horizontal stripes [39], triangulated graphs [95], concentric rings [275], and localized patches [20].

Another, complementary, taxonomy (Table 4.1) for the re-identification algorithms distinguishes the class of the *single-shot* approaches, focusing on associating pairs of images, each containing one instance of an individual, from the class of *multiple-shot* methods. The latter employs multiple images of the same person as probe or gallery elements. The assumption of the multi-shot methods is that individuals are tracked so that it is possible to gather lot of images. The hope is that the system will obtain a set of images that vary in terms of resolution, partial occlusions, illumination, poses, etc. In this way, we can build a robust signature of each individual suited for re-identification.

These four paradigms of re-identification give rise to the taxonomy reported in Table 4.1. Looking at the table, it is worth noting that direct single-shot approaches represent the case where the least information is employed. For each individual, we have a single image, whose features are independently matched against hundreds of candidates. The learning-based multi-shot approaches, instead, deal with the highest amount of information.

The proposed approaches lie in the class of the direct methods because we want to avoid to train classifier often. In general, learning-based approaches produce higher performances than the direct approaches. However, how stated before, they are not truly suited for a practical usage, in surveillance scenarios. Moreover SDALF is versatile, working both in the single and in the multi-shot modality, while (A)HPE works only when a certain number of images are available because it relies on the epitomic description.

SDALF and (A)HPE differ from the previous works for the several reasons: 1) Unlike [39] and [95], we do not rigidly link features to parts of the human structure, which is not reliable at low resolutions. 2) We do not employ discriminative learning techniques as in [175], that have to be re-trained each time a novel subject appears. In particular, (A)HPE does not simply accumulate local features with heuristics, as [104] and SDALF, but it keeps recurrent local aspects by analyzing the epitome resulting from the images of several person, that may reappear with higher probability in novel instances of the person.

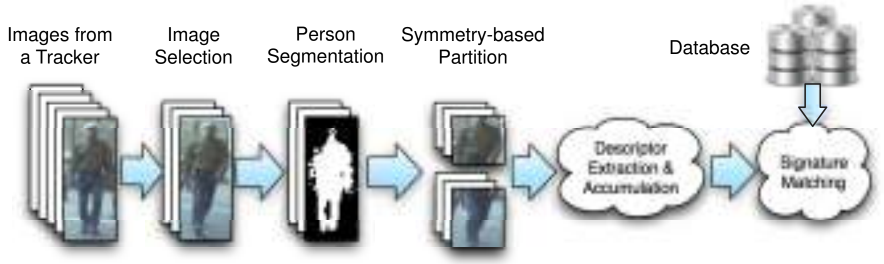


**Fig. 4.2.** Person re-identification and reacquisition pipeline. First images are gathered from tracking, then redundant images and the background pixels are discarded. The images are partitioned exploiting symmetries and asymmetries of the human body. For each part, features are extracted and accumulated over time building the signature of that person. The matching phase searches in the database to find the most similar signature.

## 4.3 Symmetry-driven Accumulation of Local Features (SDALF)

As discussed in the previous section, we assume to have a set of trackers that estimate the trajectories of each person in the different (non-)overlapped camera views as depicted in Figure 4.1. For each individual a set of bounding boxes can be obtained (from one or more consecutive frames). SDALF analyzes this images to build a signature and perform matching for recognizing individuals in a database of already-monitored individuals. The proposed re-identification pipeline of SDALF consists of six phases as depicted in Figure 4.2:

1. *Images Gathering* aggregates images given by the trajectories of the individuals and their bounding boxes.
2. *Image Selection* selects a small set of representative images, when the number of images is very high (*e.g.*, in tracking) in order to discard redundant information. [Section 4.3.1]
3. *Person Segmentation* separates the pixels of the individual (foreground) from the rest of the image (background) that usually "distracts" the re-identification. [Section 4.3.2]
4. *Symmetry-based Silhouette Partition* detects perceptually salient body regions exploiting symmetry and asymmetry principles. [Section 4.3.3]
5. *Descriptor Extraction and Accumulation* composes the signature as an ensemble of global or local features extracted from each body part and from different frames. [Section 4.3.4 and 4.6]
6. *Signature Matching* minimize a certain similarity score between the probe signature and a set of signatures collected in a database (gallery set). [Section 4.6.2]

The nature of this process is slightly different depending on if we have one or more images, that is, single- or multiple-shot case, respectively. In the following, each step is described and analyzed in details, focusing on the differences between single-shot and multi-shot modality.

### 4.3.1 Image Selection

Since there is a temporal correlation between images of each tracked individual, redundancy is expected. It is discarded by applying the unsupervised Gaussian clustering method [84] with the automatic selection of the number of clusters. Hue Saturation Value (HSV) histogram is used as feature for clustering, in order to capture appearance similarities. Then, $N_k$ images ($k$ stays for the $k$-th person) are randomly chosen from each cluster of each person, building the set $\mathbf{X}^k = \{X_n^k\}_{n=1}^{N_k}$. Experimentally, we found that clusters with low number of elements (= 3 in our experiments) usually contain outliers, such as occlusions or partial views of the person, thus these cluster are discarded. It is worth noting that the clusters the method automatically selects can still contain occlusions and bad images, hard for the re-identification task. The feature and the clustering technique we chose are simple and give a very rough result, but we tested experimentally that this is enough to obtain good results and in order to have a fast image selection method.

### 4.3.2 Person Segmentation

The aim of this phase is to separate the genuine body appearance from the rest of the scene. A lot of re-identification methods do not perform this step, that turns out to be an essential step in this problem. Since the available datasets are not properly built for the re-identification task, most of the images of the same individuals have very similar background. Thus, it seems that some methods exploit the background information to make easier the re-identification task. In other words, they exploits the background appearance to distinguish between different individuals, because different individuals have different backgrounds. However, the use of this type of context to perform re-identification is not proper for general purposes.

Person segmentation allows the descriptor to focus solely on the individual, disregarding the context in which she/he is immersed. We suppose that in a real scenario, a person can be captures at completely different locations, like the arrival hall of an airport, and in the parking lot. In the case of a sequence of consecutive images, the object/scene classification may be operated by a whatsoever background subtraction strategy. In the case of a single image, the separation is performed by Stel Component Analysis (SCA) [119].

SCA lies on the notion of "structure element" (stel), which can be intended as an image portion (often discontinuous) whose topology is consistent over an image class. This means that in a set of given objects (faces or pedestrian images), a stel individuates the same part over all the instances (e.g., the hair in a set of faces, the body in a set of images containing single pedestrians). In other words, an image can be seen as a segmentation, where each segment is a stel. SCA enriches the stel concept as it captures the common structure of an image class by blending together multiple stels: it assumes that each pixel measurement $x_i$, with its 2D coordinate $i$, has an associated discrete variable $s_i$, which takes a label from the set $\{1, \dots, S\}$. Such a labeling is generated from $K$ stel priors $p_k(s_i)$, which capture the common structure of the set of images (see Fig. 4.3 for a representative example). The model detects the image self-similarity within a segment: the pixels with the same label $s$ are expected to follow a tight distribution over the image measurements.

Instead of the local appearance similarity, the model insists on consistent segmentation via the stel prior. Each component $k$ represents a characteristic (pose or spatial configuration) of the object class at hand, and other poses are obtained through blending these components. We set $S = 2$ (i.e., foreground/background) and $K = 2$, modeling the distribution over the image measurements as a mixture of Gaussians as we want to capture segments with multiple color modes within them. The value for these components has been chosen experimentally. SCA has been learnt beforehand on a person database not considering the experimental data, and the segmentation over new samples consists in a fast inference. Each Expectation-Maximization iteration of the inference algorithm takes in average 18 milliseconds[1] when dealing with images of size 48 × 128. In our experiments, we set the number of iterations to 100.

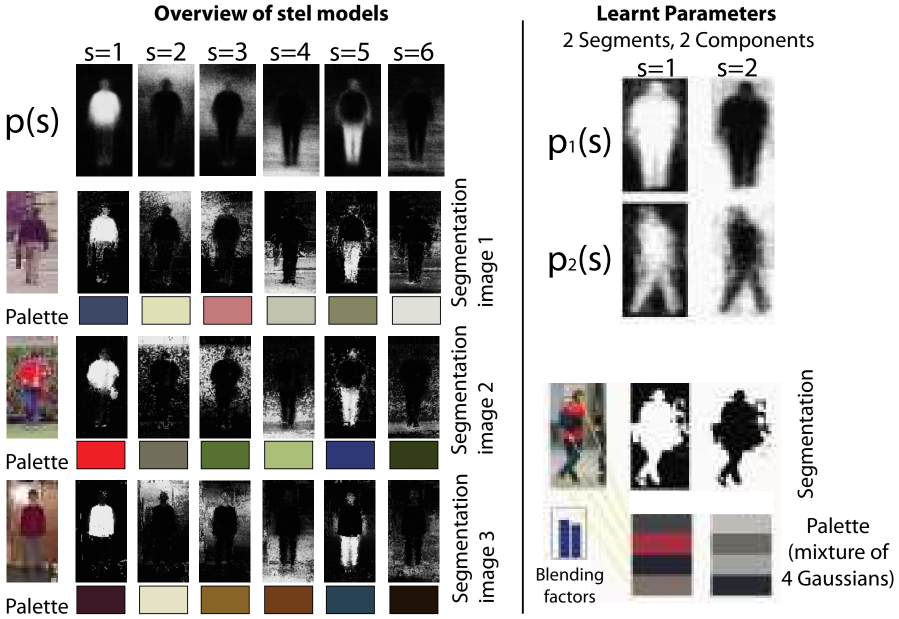


**Fig. 4.3.** Stel Component Analysis for image segmentation

### 4.3.3 Symmetry-based Silhouette Partition

The goal of this phase is to partition the human body into salient parts, exploiting asymmetry and symmetry principles. Considering a pedestrian acquired at very low resolution (see some examples in decreasing resolutions in Fig. 4.4), it is easy to note that the most distinguishable parts are three: head, torso and legs. Focusing on such parts is thus reasonable, and their detection can be exploited observing

[1] We used the authors' MATLAB code [119] on a quad-core Intel Xeon E5440, 2.83 GHz with 4 GB of RAM.

natural asymmetry properties in the human appearance. In addition, the relevance of head, torso and legs as salient regions for human characterization also emerged from the boosting approach proposed by [102].

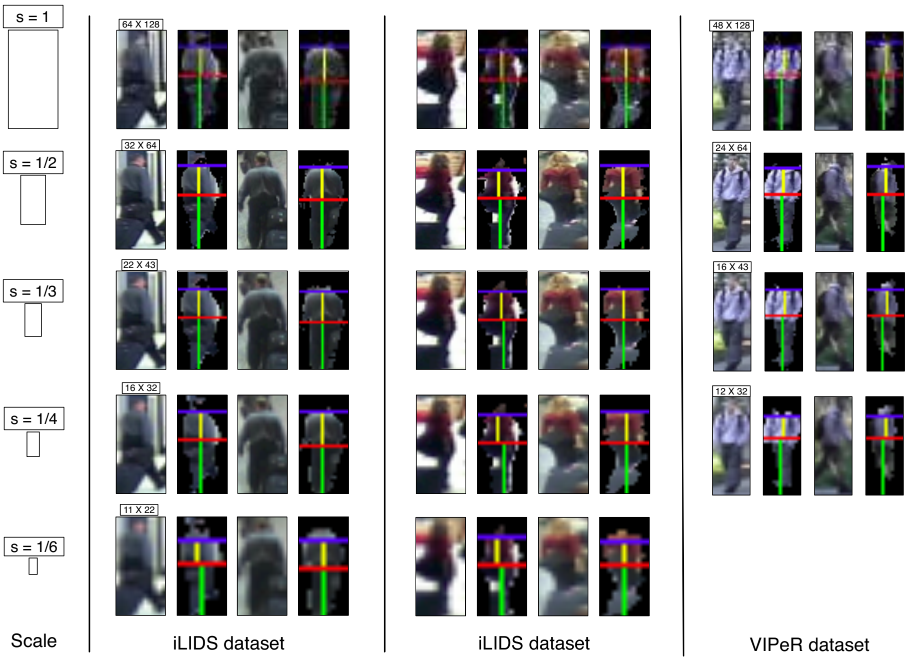


**Fig. 4.4.** Images of individuals at different resolutions (from 64 × 128 to 11 × 22) and examples of foreground segmentation and symmetry-based partitions.

Let us first define two basic operators. The first is the *chromatic bilateral operator*:

$$C(i, \delta) \propto \sum_{B_{[i-\delta,i+\delta]}} d^2 \left(p_i, \hat{p}_i\right) \tag{4.1}$$

where $d(\cdot, \cdot)$ is the Euclidean distance, evaluated between HSV pixel values $p_i, \hat{p}_i$, located symmetrically with respect to the horizontal axis at height $i$. This distance is summed up over $B_{[i-\delta,i+\delta]}$, i.e. the foreground region (i.e., that segmented by the object segmentation phase) lying in the box of width $J$ and vertical extension $2\delta + 1$ around $i$ (see Fig. 4.5). We fix $\delta = I/4$, proportional to the image height, so that scale independency can be achieved.

The second one is the *spatial covering operator*, that calculates the difference of foreground areas for two regions:

$$S(i, \delta) = \frac{1}{J\delta} \left| A\left(B_{[i-\delta,i]}\right) - A\left(B_{[i,i+\delta]}\right) \right| , \tag{4.2}$$

where $A\left(B_{[i-\delta,i]}\right)$, similarly as above, is the foreground area in the box of width $J$ and vertical extension $[i - \delta, i]$.

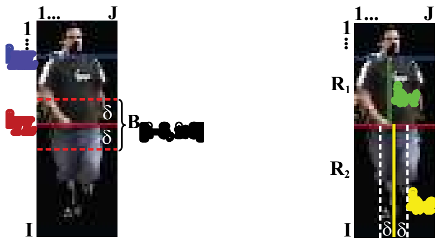


**Fig. 4.5.** Symmetry-based Silhouette Partition. On the top row, overview of the method: first the asymmetrical axis $i_{TL}$ is extracted, then $i_{HT}$; afterwards, for each $R_k$, $k = \{1, 2\}$ region the symmetrical axis $j_{LRk}$ are computed.

Combining opportunely $C$ and $S$ gives the axes of symmetry and asymmetry. The main $x$-axis of asymmetry is located at height $i_{TL}$:

$$i_{TL} = \underset{i}{\operatorname{argmin}}\ (1 - C(i, \delta)) + S(i, \delta), \tag{4.3}$$

i.e., we look for the $x$-axis that separates regions with strongly different appearance and similar area. The values of $C$ are normalized by the numbers of pixels in the region $B_{[i-\delta,i+\delta]}$. The search for $i_{TL}$ holds in the interval $[\delta,\ I - \delta]$: $i_{TL}$ usually separates the two biggest body portions characterized by different colors (corresponding to t-shirt/pants or suit/legs, for example).

The other $x$-axis of asymmetry is positioned at height $i_{HT}$, obtained as:

$$i_{HT} = \underset{i}{\operatorname{argmin}} \left(-S(i, \delta)\right). \tag{4.4}$$

This separates regions that strongly differ in area and places $i_{HT}$ between head and shoulders. The search for $i_{HT}$ is limited in the interval $[\delta, i_{TL} - \delta]$.

The values $i_{HT}$ and $i_{TL}$ isolate three regions $R_k$, $k = \{0, 1, 2\}$, approximately corresponding to head, body and legs, respectively (see Fig. 4.5). The head part $R_0$ is discarded, because it often consists in few pixels, carrying very low informative content.

At this point, for each part $R_k$, $k = \{1, 2\}$, a (vertical) symmetry axis is estimated, in order to individuate the areas that most probably belong to the human body, i.e., pixels near the symmetry axis. In this way, the risk of considering background clutter is minimized.

On both $R_1$ and $R_2$, the $y$-axis of symmetry is estimated in $j_{LRk}$, $(k = 1, 2)$, obtained using the following operator:

$$j_{LRk} = \underset{j}{\operatorname{argmin}}\ C(j, \delta) + S(j, \delta). \tag{4.5}$$

This time, $C$ is evaluated on the foreground region of size the height of $R_k$ and width $\delta$ (see Fig. 4.5). We look for regions with similar appearance and area. In this case, $\delta$ is proportional to the image width, and it is fixed to $J/4$.

In Fig. 4.4, different individuals are taken in different shots. As one can observe, our subdivision segregates correspondent portions independently on the assumed pose and the adopted resolution.

### 4.3.4 Accumulation of Local Features

Once the asymmetry/symmetry axes have been set, different features are extracted from the parts $R_1$ and $R_2$ (torso and legs, respectively). The goal is to distill as much complementary aspects as possible in order to encode heterogeneous information, so capturing distinctive characteristics of the individuals. Each feature is extracted by taking into account its distance with respect to the $j_{LRk}$ axes. The basic idea is that locations far from the symmetry axis belong to the background with higher probability. Therefore, features coming from that areas have to be a) weighted accordingly or b) discarded. Depending on the considered features, one of these two mechanisms will be applied.

There are many possible cues useful for a fine visual characterization. Considering the previous literature in human appearance modeling, features may be grouped by considering the kind of information to focus on, that is, chromatic (histograms), region-based (blobs), and edge-based (contours, textures) information. Here, we consider a feature for each aspect, showing later on their importance (see Fig. 4.6(c-e) for a qualitative analysis of the feature for the SDALF descriptor).

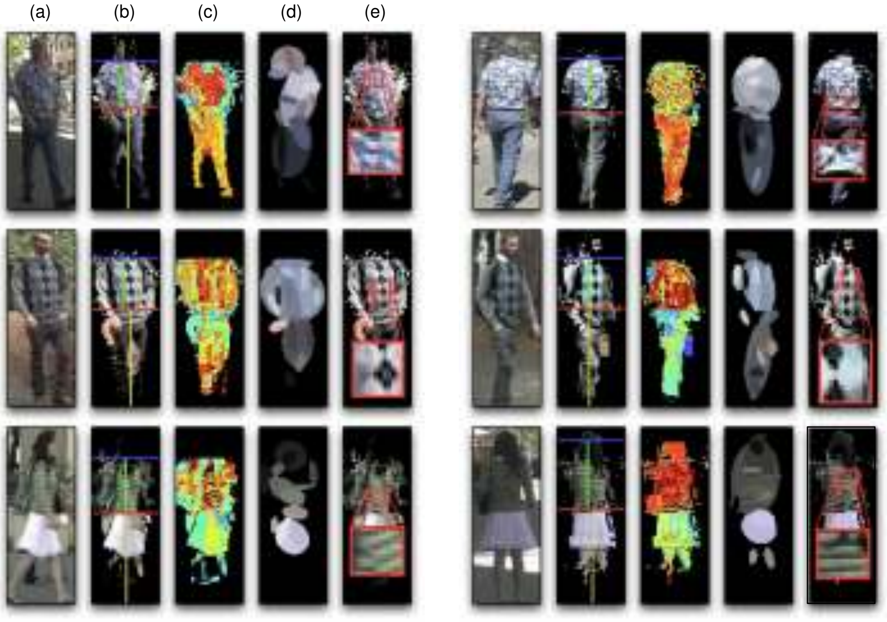


**Fig. 4.6.** Sketch of the SDALF descriptor for single-shot modality. (a) Given an image or a set of images, (b) SDALF localizes meaningful body parts. Then, complementary aspects of the human body appearance are extracted: (c) weighted HSV histogram, represented here by its weighted back-projection (brighter pixels mean a more important color), (d) Maximally Stable Color Regions [85] and (e) Recurrent Highly Structured Patches. The objective is to correctly match SDALF descriptors of the same person (first column vs sixth column).

### *Weighted Color Histograms*

The chromatic content of each part of the pedestrian is encoded by color histograms. We evaluate different color spaces, namely, HSV, RGB, normalized RGB (where each channel is normalized by the sum of all the channels), per-channel normalized RGB [20], CIELAB. Among these, HSV has shown to be superior and also allows a intuitive quantization against different environmental illumination conditions and camera acquisition settings.

Therefore, we build *weighted histograms*, so taking into consideration the distance to $j_{LRk}$ axes. In particular, each pixel is weighted by a one-dimensional Gaussian kernel $\mathcal{N}(\mu, \sigma)$, where $\mu$ is the $y$-coordinate of $j_{LRk}$, and $\sigma$ is a priori set to $J/4$. The nearer a pixel to $j_{LRk}$, the more important. In the single-shot case, a single histogram for each part is built. Instead, in the multiple-shot case, with $M$ instances, $M$ histograms for each part are considered. Then, the matching policy will handle these multiple histograms properly (see Section 4.4).

### *Maximally Stable Color Regions (MSCR)*

The MSCR operator[2] [85] detects a set of blob regions by looking at successive steps of an agglomerative clustering of image pixels. Each step clusters neighboring pixels with similar color, considering a threshold that represents the maximal chromatic distance between colors. Those maximal regions that are stable over a range of steps constitute the maximally stable color regions of the image. The detected regions are then described by their area, centroid, second moment matrix and average RGB color, forming 9-dimensional patterns. These features exhibit desirable properties for matching: covariance to adjacency preserving transformations and invariance to scale changes and affine transformations of image color intensities. Moreover, they show high repeatability, i.e., given two views of an object, MSCRs are likely to occur in the same correspondent locations.

In the single-shot case, we extract MSCRs separately from each part of the pedestrian. In order to discard outliers, we select only MSCRs that lie inside the foreground regions. In the multiple-shot case, we opportunely accumulate the MSCRs coming from the different images by employing a Gaussian clustering procedure [84], which automatically selects the number of components. Clustering is carried out using the 5-dimensional MSCR sub-pattern composed by the centroid and the average RGB color of each blob. We cluster the blobs similar in appearance and position, since they yield redundant information. The contribution of this clustering operation is twofold: i) it captures only the relevant information, and ii) it keeps low the computational cost of the matching process, when the clustering results are used. The final descriptor is built by a set of 4-dimensional MSCR sub-pattern composed by the $y$ coordinate and the average RGB color of each blob. Please note that $x$ coordinates are discarded because they are strongly dependent on the pose and viewpoint variation.

[2] We used the author's implementation, downloadable at `http://www2.cvl.isy.liu.se/~perfo/software/`.

*Recurrent High-Structured Patches (RHSP)*

We design this feature taking inspiration from the image epitome [118]. The idea is to extract image patches that are highly recurrent in the human body figure (see Fig. 4.7). Differently from the epitome, we want to take into account patches that are 1) informative (in an information theoretic sense, i.e., carrying out high entropy values), and 2) that can be affected by rigid transformations. The first constraint selects only those patches with strong edges, such as textures. The second requirement takes into account that the human body is a 3D entity whose parts may be captured with distortions, depending on the pose. For simplicity, we modeled the human body as a vertical cylinder. In these conditions, the RHSP generation consists in three phases.

The first step consists in the random extraction of patches $p$ of size $J/6 \times I/6$, independently on each foreground body part of the pedestrian. In order to take the vertical symmetry into consideration, we mainly sample the patches around the $j_{LRk}$ axes, exploiting the Gaussian kernel used for the color histograms computation. In order to focus on informative patches, we operate a thresholding on the entropy values of the patches, pruning away patches with low structural information (e.g., uniformly colored). This entropy is computed as the sum $H_p$ of the pixel entropy of each RGB channel. We choose those patches with $H_p$ higher than a fixed threshold $\tau_H$ ( $= 13$ in all our experiments). The second step applies a set of transformations $T_i$, $i = 1, 2, \ldots, N_T$ on the generic patch $p$, for all the sampled $p$'s in order to check their invariance to (small) body rotations, i.e., considering that the camera may capture the one's front, back or side, and supposing the camera is at the face's height. We thus generate a set of $N_T$ *simulated* patches $p_i$, gathering an enlarged set $\hat{p} = \{p_1, \ldots, p_{N_T}, p\}$.

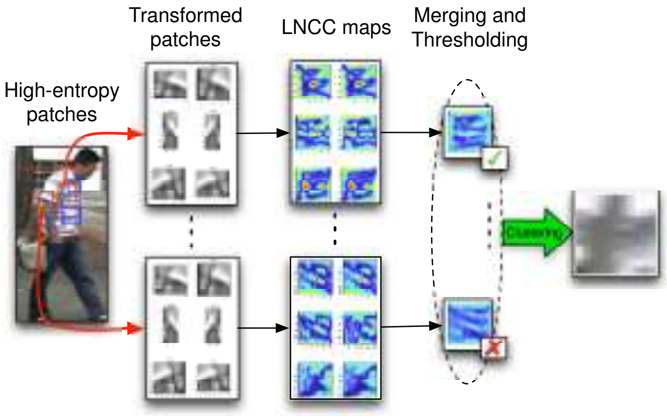


**Fig. 4.7.** Recurrent high-structured patches extraction. The final result of this process is a set of patches (in this case only one) characterizing each body part of the pedestrian.

In the third and final phase, we investigate how much recurrent a patch is. We evaluate the Local Normalized Cross-Correlation (LNCC) of each patch in $\hat{p}$ with respect to the original image. All the $N_T + 1$ LNCC maps are then summed together forming an average map. Averaging again over the elements of the map indicates how much a patch, and its transformed versions, is present in the image. Thresholding this value ($\tau_\mu = 0.4$) does select the RHSP patches.

Given a set of RHSPs for each region $R_1$ and $R_2$, the descriptor consists again of an HSV histogram of these patches. We have tried experimentally to use local binary pattern descriptor, that describe the texture, but it turned out to be less robust than color histograms. We think that the main reason for this is that the extracted patches have very low resolution, therefore a color description capture more invariance. The single-shot and the multiple-shot methods are similar, with the only difference that in the multi-shot case the candidate RHSP descriptors are accumulated over different frames.

Please note that, even if we have several thresholds that regulate the feature extraction, they have been fixed once, and left unchanged in all the experiments. The best values have been selected by cross-validation on a half of a re-identification dataset (in our experiments, VIPeR dataset).

## 4.4 Signature Matching

In this section, we illustrate how the different features are employed as a single signature for re-identification. In a general re-identification problem we have two sets of pedestrian signatures: a gallery set $A$ and a probe set $B$. Re-identification consists in associating each person of $B$ to the corresponding person of $A$. We denote $\mathbf{P}^A$ and $\mathbf{P}^B$, the signature of an individual in the set $A$ and in the set $B$, respectively. The association mechanism depends on how the two sets are organized, more specifically, on how many pictures are present for each individual. This gives rise to three matching philosophies: 1) *single-shot vs. single-shot* (SvsS), if each image in a set represents a different individual; 2) *multiple-shot vs. single-shot* (MvsS), if each image in $B$ represents a different individual, while in $A$ each person is portrayed in different images, or *instances*; 3) *multiple-shot vs. multiple-shot* (MvsM), if both $A$ and $B$ contain multiple instances per individual.

In general, we can define the re-identification issue as a maximum log-likelihood estimation problem. More in details, given a probe $B$ the correct matching is carried out by:

$$A^* = \arg\max_A \left(\log P(\mathbf{P}^A|\mathbf{P}^B)\right) = \arg\min_A \left(d(\mathbf{P}^A, \mathbf{P}^B)\right) \tag{4.6}$$

where the equality is valid because we define $P(\mathbf{P}^A|\mathbf{P}^B)$ in Gibbs form $P(\mathbf{P}^A|\mathbf{P}^B) = e^{-d(\mathbf{P}^A,\mathbf{P}^B)}$ and $d(\mathbf{P}^A, \mathbf{P}^B)$ measures the distance between two descriptors. In the next section, we will see how we defined the distance $d$ for SDALF.

### 4.4.1 SDALF Matching

The *SDALF matching distance* $d$ is defined as a convex combination of the local features:

$$d(\mathbf{P}^A, \mathbf{P}^B) = \beta_{\mathrm{WH}} \cdot d_{\mathrm{WH}}(\mathrm{WH}(\mathbf{P}^A), \mathrm{WH}(\mathbf{P}^B)) + \tag{4.7}$$

$$\beta_{\mathrm{MSCR}} \cdot d_{\mathrm{MSCR}}(\mathrm{MSCR}(\mathbf{P}^A), \mathrm{MSCR}(\mathbf{P}^B)) + \tag{4.8}$$

$$\beta_{\mathrm{RHSP}} \cdot d_{\mathrm{RHSP}}(\mathrm{RHSP}(\mathbf{P}^A), \mathrm{RHSP}(\mathbf{P}^B)) \tag{4.9}$$

where the WH($\cdot$), MSCR($\cdot$), and RHSP($\cdot$) are the weighted histograms, MSCR, and Recurrent High-Structured Patch descriptors, respectively, and $\beta$s are normalized weights.

The distance $d_{\mathrm{WH}}$ considers the weighted color histograms. In the SvsS case, the HSV histograms of each part are concatenated channel by channel, then normalized, and finally compared via Bhattacharyya distance [121]. Under the MvsM and MvsS policies, we compare each possible pair of histograms contained in the different signatures, keeping the lowest distance.

For $d_{\mathrm{MSCR}}$, in the SvsS case, we estimate the minimum distance of each MSCR element $b$ in $\mathbf{P}^B$ to each element $a$ in $\mathbf{P}^A$. This distance is defined by two components: $d_y^{ab}$, that compares the $y$ component of the MSCR centroids; the $x$ component is ignored, in order to be invariant with respect to body rotations. The second component is $d_c^{ab}$, that compares the MSCR color. In both cases, the comparison is carried out using the Euclidean distance.

The two components are combined as:

$$d_{\mathrm{MSCR}} = \sum_{b \in \mathbf{P}^B} \min_{a \in \mathbf{P}^A} \gamma \cdot d_y^{ab} + (1 - \gamma) \cdot d_c^{ab} \tag{4.10}$$

where $\gamma$ takes values between 0 and 1. In the multi-shot cases, the set $\mathbf{P}^A$ of Eq. 4.10 becomes a subset of blobs contained in the most similar cluster to the MSCR element $b$.

The distance $d_{\mathrm{RHSP}}$ is obtained by selecting the best pair of RHSP, one in $\mathbf{P}^A$ and one in $\mathbf{P}^B$, and evaluating the minimum Bhattacharyya distance among the RHSP's HSV histograms. This is done independently for each body part (excluding the head), summing up all the distances achieved, and then normalizing for the number of pairs.

In our experiments, we fix the values of the parameters as follows: $\beta_{\mathrm{WH}} = 0.4$, $\beta_{\mathrm{MSCR}} = 0.4$, $\beta_{\mathrm{RHSP}} = 0.2$ and $\gamma = 0.4$. These values are estimated by cross validating over the first 100 image pairs of the VIPeR dataset, and left unchanged for all the experiments.

### 4.4.2 Detecting new instances

The literature of re-identification does not take into account the case where an individual $\mathbf{P}^B$ is not already in the gallery set. In a real re-identification setting, this is a very frequent scenario, where new people enters the scene for the first time ever.

We address this issue by observing the distribution of the distances of correct matches and the distances of wrong matches. Experimentally, we have found out that these distances follow the bimodal distribution of Fig. 4.8, where the correct matching distances and the wrong matching distances are depicted by the green (on the left) and the red histogram (on the right), respectively. By simply fitting two

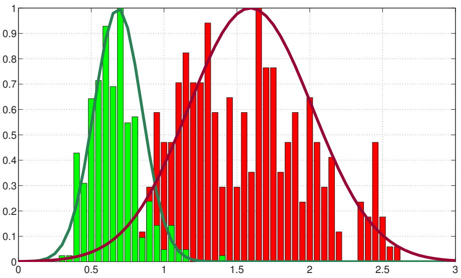


**Fig. 4.8.** Bimodal distribution of distances. The correct matching distances and the wrong matching distances are depicted by the green and the red histogram, respectively. These curves have been computed for the ETHZ, MvsM N=2 experiment, discussed in Section 4.7.

Gaussian distributions on the distances data, we are able to distinguish between correct matches and wrong matches. This means that given the minimum distance $d(\mathbf{P}^{A^*}, \mathbf{P}^B)$ between the best matching $\mathbf{P}^{A^*}$ and $\mathbf{P}^B$, if $d(\mathbf{P}^{A^*}, \mathbf{P}^B)$ is associated to the mode of "wrong" distances (Fig. 4.8 on the left), the individual is not in the gallery set. If $d(\mathbf{P}^{A^*}, \mathbf{P}^B)$ is associated to the other mode (Fig. 4.8 on the right), re-identification is performed. Instead of manually choosing a threshold, that may have to be changed for different scenarios, the likelihood ratio of the two Gaussian can be exploited. We estimate the parameters $\mu_1, \sigma_1, \mu_2, \sigma_2$ of the two Gaussians ($\mathcal{N}(\mu_1, \sigma_1)$ for the mode of "correct" distances and $\mathcal{N}(\mu_2, \sigma_2)$ for the mode of the "wrong" distances) in a training phase, shown in Fig. 4.8. At testing time, given a distance $d$ if $\frac{\mathcal{N}(d;\mu_1,\sigma_1)}{\mathcal{N}(d;\mu_2,\sigma_2)} >= 1$ there is re-identification, otherwise we identify a new individual. Alternatively, we can use the log-likelihood ratio that is usually numerically more stable, $\log\mathcal{N}(d;\mu_1,\sigma_1) - \log\mathcal{N}(d;\mu_2,\sigma_2) >= 0$, where $\log\mathcal{N}(\mu,\sigma)$ is the log formulation of the Gaussian distribution, without performing the exponentiation, that is, $\log\mathcal{N}(x;\mu,\sigma) = -\frac{(x-\mu)}{2\sigma^2} - \sqrt{2\pi\sigma^2}$.

## 4.5 SDALF as Appearance Descriptor for Tracking

The probabilistic framework used here is the tracking-by-detection algorithm based on particle filtering [50,178] described in Section 2.2.2. In our multi-target scenario, each individual is tracked by an independent particle filter for simplicity, *i.e.*, occlusions are not modeled. Therefore, each target has its own proposal distribution that depends on an associated detection. Each target is associated to a detection that is nearest in terms of spatial distance and appearance similarity. This means that at each step we perform data association between the measurements at current time (the detections) and the state estimates at previous time (the estimated bounding boxes) based on nearest neighbor.

For person detection, the already-trained person detector proposed in [80] has been used. For generating new tracks, weak tracks are kept in memory, and it is checked whether or not they are supported continuously by a certain amount of detections. If this happens, the track is initialized as in [43]. It is worth noting that our purpose is to highlight the quality of our appearance model, and not to propose a novel tracking algorithm. For this reason, the tracking framework is standard and basic. One can use other more advanced methods, such as the HJS filter presented in Section 2.2.1 or other techniques discussed in Section 2.1.

Given the assumptions that our method relies on, let us discuss of the main contribution of this section. We propose a novel observation model $p(\mathbf{y}_t|\mathbf{x}_t^{(n)})$ that can be easily be embedded in any particle filter. The novelty is mainly in using the SDALF descriptor as object representation. We define the observation model considering the probabilistic value of Eq. 4.6 instead of maximizing: $p(\mathbf{y}_t|\mathbf{x}_t^{(n)}) = P(\mathbf{P}^A|\mathbf{P}^B)$, where in this case $\mathbf{P}^B$ is the object template made by SDALF descriptors, and $\mathbf{P}^A$ is the current hypothesis $x_t^{(n)}$. In this case, we do not optimize Eq. 4.6, but the full probability distribution over the hypotheses is kept in order to embed it into the probabilistic framework of particle filtering.

In order to fit the descriptor into the tracking problem we need to make some simplifications. First of all, since the descriptor has to be extracted for each hypothesis $\mathbf{x}_t^{(n)}$, it should be reasonably efficient. In our current implementation, the computation of RHSP for each particle is not feasible as the transformations $T_i$ performed on the original patches to make the descriptor invariant to rigid transformations constitute a too high burden. A simple solution could be to consider only a low number of transformations, but we found experimentally that in that case we obtain worse results than removing directly RHSP from the descriptor. Therefore, the RHSP is drop out from SDALF for tracking.

More formally, the observation model becomes:

$$p(\mathbf{y}_t|\mathbf{x}_t^{(n)}) \propto e^{-D(\mathbf{x}_t^{(n)}(\mathbf{y}_t),\tau_t)} \tag{4.11}$$

where

$$\begin{aligned} D(\mathbf{x}_t^{(n)}(\mathbf{y}_t),\tau_t) =& \beta_{\mathrm{WH}} \cdot d_{\mathrm{WH}}(\mathrm{WH}(\mathbf{x}_t^{(n)}(\mathbf{y}_t)), \mathrm{WH}(\tau_t)) + \\ & \beta_{\mathrm{MSCR}} \cdot d_{\mathrm{MSCR}}(\mathrm{MSCR}(\mathbf{x}_t^{(n)}(\mathbf{y}_t)), \mathrm{MSCR}(\tau_t)) \end{aligned}$$

where $\mathbf{x}_t^{(n)}(\mathbf{y}_t)$ is the patch extracted from the image $\mathbf{y}_t$ given the bounding box $\mathbf{x}_t^{(n)}$, and $\tau_t$ is the template image of the object. During tracking, the object template has to be updated in order to model the different aspects of the captured object (for example, due to different poses). $M = 3$ images are randomly selected from a temporal window with a fixed length $L$ at each time step to build $\tau_t$, in order to balance the number of images employed for building the model and the computational effort required. Random selection of images to build the template is not always the best choise because sometimes the model can drift away to the background. Also in our experiments, we saw this issue happends. Many techniques can be used to avoid or at least slow down the drifting process, such as, the selection of good templates with P-N learning [122].

The computation of the observation model of Eq. 4.11 consists in evaluating the distances of the hypotheses $\{\mathbf{x}_t^{(n)}\}$ (single images) against $\tau_t$ ($M$ images), as dictated by the MvsS strategy of the re-identification task. Using the re-identification paradigm, we have a gallery composed by a set of $M$ images, and a bunch of probe images.

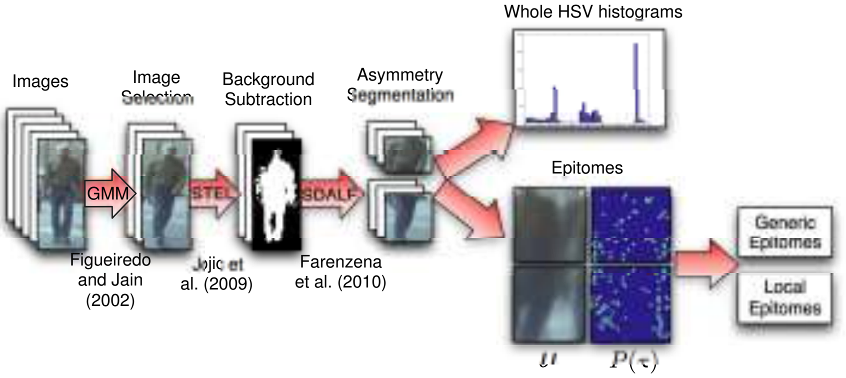


**Fig. 4.9.** Overview of the proposed approach.

## 4.6 Chromatic and Epitomic Analyses

In this section, we discuss how to "hack" the pipeline proposed in SDALF in order to accumulate more properly features over time. The overview of the proposed approach is shown in Fig. 4.9. The first four points are exactly the same steps we have seen in SDALF. What is different here is the descriptors we choose and how they are accumulated other time (last two steps of Fig. 4.9). The main idea is to perform both chromatic and *epitomic* analyses of the images in order to extract a robust signature of each individual. We propose two variant of the method: Histogram Plus Epitome (HPE) and Asymmetry-based HPE (AHPE). The only difference is that HPE does not have symmetry-based silhouette partition phase while AHPE does have it, and thus HPE operates on the whole image.

### 4.6.1 Histogram Plus Epitome (HPE)

We define the HPE descriptor as composition of three features extracted from each $\mathbf{X}^k$: a chromatic global feature, that is, a color histogram; and two local epitome-based features, that capture the presence of recurrent local patterns. Moreover, those features can be extracted from parts of the person (Section 4.6.3).

*Color histogram*

As global appearance feature we use the Hue Saturation Value (HSV) histogram, proven to be very effective and largely adopted in several applications [102, 224]. We encode it in a 36-dimensional feature space $[H = 16, S = 16, V = 4]$, one for each instance. Then, the global feature $\mathrm{H}(\cdot)$ is built by averaging the histograms of the multiple instances of $\mathbf{X}^k$. This makes the feature robust to illumination and pose variations, keeping the predominant chromatic information.

*Epitomic Analysis*

The main contribution of the work is that we employ the epitomic analysis by [118] to accumulate information/images over the time in order to build a multi-shot descriptor, without any assumption or heuristics. An image epitome is the result of collapsing an image or a set of images, through a generative model, into a small collage of overlapped patches embedding the essence of the textural, shape and appearance properties of the data.

A set of $P$ *ingredient* patches of fixed size[3] $I_e \times J_e$ are uniformly sampled from each image $X_n^k \in \mathbf{X}^k$, building a multi-shot set of patches $\{z_m\}_{m=1}^{N_k \times P}$. For each patch $z_m$, the generative model infers a hidden mapping variable $\tau_m(i,j)$ that maps (through translations) $z_m$ into a equally sized portion of the epitome, having $(i,j)$ as left-upper corner. The inference is possible by evaluating the variational distribution $q(\tau_m(i,j))$, that represents the probability of that mapping (see [118] for details). By mapping all patches in the epitome space and averaging them, we extract the epitome's parameters $e = \{\mu, \phi\}$, where $\mu$ is the epitome mean, *i.e.*, an image that contains similar, recurrent patches present in several instances, while $\phi$ represents the standard deviation map associated to each pixel of the epitome.

We customize the use of the epitome for the task at hand, extracting two different features from it: the *generic* epitome and the *local* epitome. The generic epitome Ge$(\cdot)$ extracts information directly from the mean $\mu$. Considering just $\mu$ is equivalent to disregarding (*i.e.*, being invariant to) small variations among the different instances' patches, usually due to small scale/pose discrepancies and illumination variations. A single HSV histogram is obtained from $\mu$ in order to have a robust appearance-based feature. Moreover, learning an epitome twice on the same data gives two similar models with a different spatial displacement. Adopting histograms cancels out such discrepancy.

On the other hand, the local epitome Le$(\cdot)$ is focused on detecting local regions in the epitome that portray highly informative recurrent ingredient patches. To this end, first, we estimate the prior probability on the transformation $P(\tau) = \frac{\sum_m q(\tau_m)}{N_k \cdot P}$ (see Fig. 4.9), that gives the probability that the patch in the epitome having $(i,j)$ as left-upper corner represents several ingredient patches $\{z_m\}$. Second, we rank in descending order of $P(\tau)$ all the patches in the epitome, retaining only the first $M = 40$, *i.e.*, the most recurrent ones. Then, we rank again these $M$ patches in descending order by evaluating their entropy, retaining the first $F = 10$, *i.e.*, the most informative ones[4]. We describe each *survived* patch with an HSV histogram (*i.e.*, $F$ histograms in total).

### 4.6.2 HPE Matching

Similarly to SDALF, the *HPE matching distance* $d$ is defined by combining three similarities scores (one for each feature):

$$\begin{aligned} d(\mathbf{P}^A, \mathbf{P}^B) = \beta_1 \cdot (d_c(\mathrm{H}(\mathbf{P}^A), \mathrm{H}(\mathbf{P}^B))) + \\ \beta_2 \cdot (d_c(\mathrm{Ge}(\mathbf{P}^A), \mathrm{Ge}(\mathbf{P}^B))) + \\ \beta_3 \cdot (d_e(\mathrm{Le}(\mathbf{P}^A), \mathrm{Le}(\mathbf{P}^B))) \end{aligned} \tag{4.12}$$

[3] To set the patch sizes we should fulfill the trade-off between too small patches, where the epitome converges to an histogram, and too big patches where the epitome loose its generalization properties. Experimentally, we found out that the patch area has to be 1/3 of the area of the image.

[4] $M$ and $F$'s values are set after cross-validation on a small experimental data subset.

where the $\mathrm{H}(\cdot)$, $\mathrm{Ge}(\cdot)$, and $\mathrm{Le}(\cdot)$ are the HSV histogram, the generic and the local epitome, respectively, and $\beta$s are normalized weights[5]. $d_c$ is the Bhattacharyya distance, while $d_e$ is estimated as the minimum distance of each patch $b$ in $\mathrm{Le}(\mathbf{P}^B)$ to each patch $a$ in $\mathrm{Le}(\mathbf{P}^A)$ of the local epitome, *i.e.*:

$$d_e = \frac{1}{C} \sum_{b \in \mathrm{Le}(\mathbf{P}^B)} \min_{a \in \mathrm{Le}(\mathbf{P}^A)} d_c(\mathrm{H}(a), \mathrm{H}(b)), \tag{4.13}$$

where $C$ is a normalization constant.

In terms of computational complexity, Eq. 4.6 is bounded by $\mathcal{O}(K \cdot (N^2 + F^2))$, because we have $K$ pedestrians with $N$ images each, which means $K \cdot N$ HSV histograms, $K$ $\mathrm{Ge}(\cdot)$ histograms, and $K \cdot F$ $\mathrm{Le}(\cdot)$ histograms.

### 4.6.3 Asymmetry-based HPE

Semantic segmentation of objects has been largely exploited for characterizing salient parts of a structured object in object recognition tasks. We exploit the segmentation technique presented in Section 4.3.3 that uses Gestalt theory considerations on symmetry and asymmetry to segment the human body into horizontal stripes corresponding to head, torso and legs. The main idea is that horizontal parts are asymmetric in size and in appearance. The advantage of this strategy is that individuates body parts which are dependent on the visual and positional information of the clothes, robust to pose, viewpoint variations, and low resolution (where pose estimation techniques usually fail or cannot be satisfactorily applied). The AHPE matching is defined by averaging the values of Eq. 4.12 for each part.

## 4.7 Experiments

In this section, an exhaustive analysis of the methods proposed in this chapter is presented. First, the used dataset and evaluation measurements are detailed in Section 4.7.1. In Section 4.7.2 and Section 4.7.3, SDALF and (A)HPE are evaluated against the state-of-the-art methods showing the best performance in literature. Finally, we prove that SDALF can be used also as a robust appearance descriptor for tracking applications (Section 4.7.4).

### 4.7.1 Datasets and Evaluation Measurements

In literature, five different datasets were available: VIPeR [101], iLIDS for re-id [7], ETHZ 1, 2, and 3 [2]. These datasets cover challenging aspects of the person re-identification problem, such as shape deformation, illumination changes, occlusions, image blurring, very low resolution images, *etc.* However, these dataset are not exactly built for the re-identification task. In fact, the images do not come from different cameras. For this reason we set up a new dataset named

[5] See Section 4.7.3 for a quantitative analysis of the performances when these weights vary.

CAVIAR4REID [26] to merge together video surveillance challenges like the wide range of poses and real surveillance footage in iLIDS, and the multiple images and wide range of resolutions of ETHZ. To take full advantage of these conditions, we decided to take probe and gallery images from different cameras, one image each for single-shot, $M$ for multi-shot. In this settings, we have an actual re-identification dataset. Let us describe the different datasets.

**VIPeR Dataset [8, 101].** This dataset contains two views of 632 pedestrians. Each pair is made up of images of the same pedestrian taken from different cameras, under different viewpoints, poses and lighting conditions. All images are normalized to 48 × 128 pixels. Most of the examples contains a viewpoint change of 90 degrees. Each pair is randomly split into two sets: CAM A and CAM B. It is one of the most challenging one-shot datasets currently available for pedestrian re-identification. It is not possible to use this dataset to test the multi-shot modality of our methods because it contains only two images for each individual.

**ETHZ Dataset [2, 222].** The data are captured from moving cameras in a crowded street. The challenges covered by this dataset are illumination changes, occlusions and low resolution (32 × 64 pixels). This dataset contains three sub-datasets: ETHZ1 with 83 people (4.857 images), ETHZ2 with 35 people (1.936 images), and ETHZ3 contains 28 with (1.762 images). Even if this dataset does not mirrors a genuine re-identification scenario but instead a person re-acquisition scenario (no different non-overlapping cameras are employed), it still carries important challenges not exhibited by other public dataset, as the high number of images per person.

**iLIDS for re-identification Dataset [275].** The iLIDS Multiple-Camera Tracking Scenario dataset is a public video dataset captured at a real airport arrival hall in the busy times under a multi-camera CCTV network. In [275], iLIDS for re-identification dataset has been built from iLIDS Multiple-Camera Tracking Scenario. The dataset is composed by 479 images of 119 people. The images, normalized to $64 \times 128$ pixels, derive from non-overlapping cameras, under quite large illumination changes and subject to occlusions (not present in VIPeR). However, this dataset does not fit well in a multi-shot scenario because the average number of images per person is 4, and thus some individuals have only two images. In tracking applications, it is usually possible to accumulate a higher number of instances per person (one for each frame). For this reason, we also created a modified version of the dataset, named iLIDS$_{\geq 4}$, where we selected the subset of individuals with at least 4 images. In total, iLIDS$_{\geq 4}$ contains 69 individuals.

**CAVIAR for re-identification Dataset [26, 53].** CAVIAR4REID is a new dataset that contains images of pedestrians extracted from CAVIAR dataset [1].CAVIAR dataset consists of several sequences filmed in the entrance lobby of the INRIA Labs and in a shopping centre in Lisbon. We selected the latter, because the camera of the former is located overhead. Shopping centre dataset is made up by 26 sequences recorded from two different points of view at the resolution of 384 × 288 pixels. It includes people walking alone, meeting with others, window shopping, entering and exiting shops. The ground truth has been used to extract the bounding box of each pedestrian. After a manual preprocessing of the resulting set of images, a total of 72 unique pedestrians have been identified: 50 with both the camera views and 22 with one camera view. For each pedestrian, we selected a set of 10 images for each camera view in order to maximize the variance with respect to resolution changes, light conditions, occlusions, and pose changes so as to make hard re-identification. Note that the re-identification manually performed by the human expert for the creation of the dataset was a really hard task. This highlights the hardness of the dataset for the proposed automatic re-identification methods. The reader can better appreciate this claim from the examples reported in Figure 4.10.

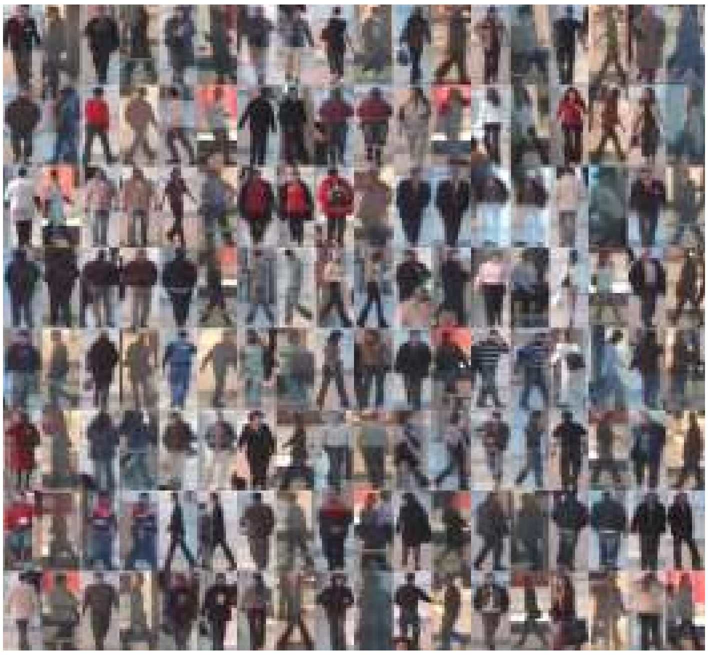


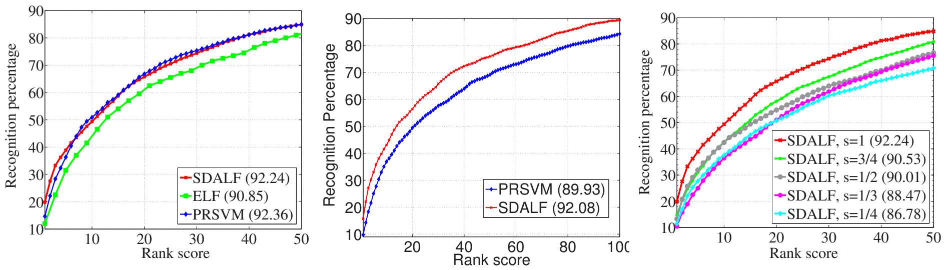
**Fig. 4.10.** Image samples from CAVIAR4REID of all the 72 pedestrians.

The main differences of CAVIAR4REID with respect to the already-existing datasets for re-identification are: 1) it has broad changes of resolution, the minimum and maximum size of the images contained on CAVIAR4REID dataset is 17 × 39 and 72 × 144, respectively. 2) Unlike ETHZ, it is extracted from a real scenario where re-identification is necessary due to the presence of multiple cameras and 3) pose variations are severe. 4) Unlike VIPeR, it contains more than one image for each view. 5) It contains the union of all the images variations of the other datasets.

**Evaluation Measures.** State-of-the-art measurements are used in order to compare the proposed methods with the others: the Cumulative Matching Characteristic (CMC) curve represents the expectation of finding the correct match in the top $n$ matches and the normalized Area Under the Curve (nAUC) is the area under the entire CMC curve normalized over the total area of the graph. nAUC gives an overall score of how well the re-identification methods do perform. We compare the proposed methods (SDALF, HPE and AHPE) with the best performances obtained so far on the available datasets: Ensemble of Localized Features (ELF) [102] and Primal-based Rank-SVM (PRSVM) [199] in VIPeR, Partial Least Squares (PLS) by [222] in ETHZ, Context-based re-id [275] and Spatial Covariance Region (SCR) [20] in iLIDS.

### 4.7.2 SDALF

In this evaluation, we consider five different datasets, VIPeR, iLIDS for re-id, ETHZ 1, 2, and 3. Each one covers different aspects and challenges for the person re-identification problem[6]. In addition, we make the task more challenging by downsampling the images up to $11 \times 22$ in order to test the methods at extremely low resolutions.



(a) 316 ped. (b) 474 ped. (c) scale, 316 ped.

**Fig. 4.11.** Performances on the VIPeR dataset in terms of CMC and nAUC (within brackets). In (a) and (b), comparative profiles of SDALF and state of the art methods (ELF [102] and PRSVM [199]) on 316 pedestrian dataset and 474 pedestrian dataset, respectively. In (c), comparison of SDALF at different scales.

Considering first VIPeR, we define Cam B as the gallery set, and Cam A as the probe set; each image of the probe set is matched with the images of the gallery. This provides a ranking for every image in the gallery with respect to the probe. Ideally rank 1 should be assigned only to the correct pair matches. The best performance so far on VIPeR dataset is obtained by PRSVM [199], following the experimental protocol of [102]. In this work, the dataset is split evenly into a training and a test set, and their algorithm, called ELF, is applied. In both algorithms a set of few random permutations are performed (5 for PRSVM, 10 for ELF), and the averaged score is kept. In order to fairly compare our results with theirs, we should know precisely the splitting assignment. Since this information

[6] A video that shows examples of SDALF descriptor and matching has been reported at `http://www.youtube.com/watch?v=3U5Aacyg-No`.

is not provided we compare the existent results with the average of the results obtained by our method for 10 different random sets of 316 pedestrians and 474 pedestrians. In Fig. 4.11, we depict a comparison among ELF, PRSVM and SDALF in terms of CMC curves. We provided also the nAUC score for each method (within brackets in the legend of the plots of Fig. 4.11). Considering the experiment on 316 pedestrians (Fig. 4.11(a)), SDALF outperforms ELF in terms of nAUC, and we obtain comparable results with respect to PRSVM. Even if PRSVM is slightly superior to SDALF, one can note that the differences between it and SDALF are negligible (less than 0.12%). This is further corroborated looking at the different philosophy underlying the PRSVM and our approach. In the former case, PRSVM uses the 316 pairs as training set, whereas in our case we act directly on the test images, operating on each single image as an independent entity. Thus, no learning phase is needed for our descriptor. In addition, it is worth noting that SDALF slightly outperforms PRSVM in the first positions of the CMC curve (rank $1-6$). This means that in a real scenario where only the first ranks are considered, our method performs better.

Fig. 4.11(b) shows a comparison between PRSVM and SDALF when dealing with a larger test dataset where a set of 474 individuals has been extracted, as done in the PRSVM paper. This reconfirms how the performances of PRSVM depend on the training set, which is now composed by 158 individuals. In this case, our approach outperforms PRSVM showing an advantage in terms of nAUC of about 2.15%. Our major erroneous matchings are due to the severe lighting changes, and to the fact that many people tend to dress in very similar ways. In these cases, additional and/or other cues are necessary, *e.g.*, considering higher resolution images, in order to grab finer image details or to consider spatio-temporal information as in our multi-shot modality.

The last analysis of this dataset consists on testing the robustness of SDALF when the image resolution decreases. We scaled the original images of the VIPeR dataset by factors $s = \{1, 3/4, 1/2, 1/3, 1/4\}$ reaching a minimum resolution of $12 \times 32$ pixels (Fig. 4.4 on the right). The results, depicted in Fig. 4.11, show that the performance decreases, as expected, but not drastically. nAUC slowly drops down from 92.24% at scale 1 to 86.78% at scale 1/4.

Now let us analyze the results on iLIDS dataset. Regarding the single-shot case, Context-based method [275] and SCR [20] produces the best performances on this dataset. We reproduce the same experimental settings of [275] in order to make a fair comparison. We randomly select one image for each pedestrian to build the gallery set, while the others form the probe set. Then, the matching between probe and gallery set is estimated. For each image in the probe set the position of the correct match is obtained. The whole procedure is repeated 10 times, and the average CMC curves is displayed in Fig. 4.12.

SDALF outperforms the Context-based method [275] without using any additional information about the context (Fig. 4.12(a)) even using images at lower resolution (Fig. 4.12(b)). The experiments of Fig. 4.12(b) show SDALF when scaling factors are $s = \{1, 3/4, 1/2,1/3, 1/4, 1/6\}$ with respect to the original size of the images, reaching a minimum resolution of $11 \times 22$ pixels. Fig. 4.12(a) shows that we get lower performances with respect to SCR [20]. Unfortunately, it has been applied solely to the iLIDS dataset; therefore, an extensive comparison on

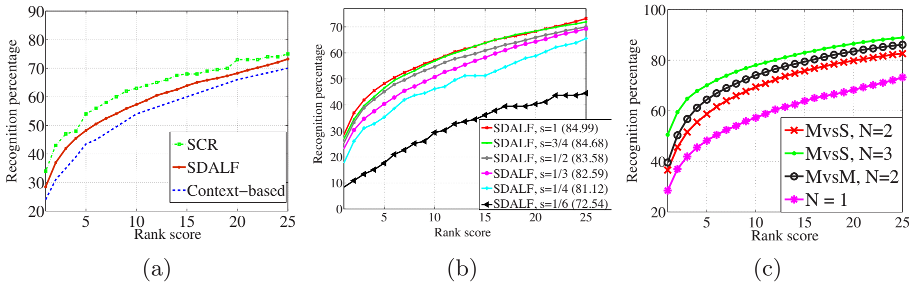


**Fig. 4.12.** Performances on iLIDS dataset. (a) CMC curves comparing Context-based re-id [275], SCR [20] and single-shot SDALF. (b) Analysis of SDALF performances at different resolution. (c) CMC curves for MvsS and MvsM cases varying the average number of images $N$ for each pedestrian. For reference, we put also the single-shot case ($N = 1$). In accordance with what reported by [275], only the first 25 ranking positions of the CMC curves are displayed.

the other datasets, each of them presenting diverse issues for re-identification, it is not possible. In particular, an interesting challenge would be that of working on extremely low resolutions, as in the ETHZ benchmarks. In [20] covariances of features are computed on localized patches. At a very low resolution this would mean computing second order statistics on very few values, that could be uninformative and subjected to dimensionality issues.

Concerning the multiple-shot case, we run experiments on both MvsS and MvsM cases. In the former trial, we built a gallery set of multi-shot signatures and we matched it with a probe set of one-shot signatures. In the latter, both gallery and probe sets are made up of multi-shot signatures. In both cases, the multiple-shot signatures are built from $N$ images of the same pedestrian randomly selected. Since the dataset contains an average of about 4 images per pedestrian, we tested our algorithm with $N = \{2, 3\}$ for MvsS, and just $N = 2$ for MvsM running 100 independent trials for each case. It is worth noting that some of the pedestrians have less than 4 images, and in this case, we simply build a multi-shot signature composed by less instances. In the MvsS strategy, this applies to the gallery signature only, and in the MvsM signature, we start by decreasing the number of instances that compose the probe signature, leaving unchanged the gallery signature; once we reach just one instance for the probe signature, we start decreasing the gallery signature too. The results, depicted in Fig. 4.12(c), show that, in the MvsS case, just 2 images are enough to increment the performances of about 10% and to outperform the Context-based method [275] and SCR [20]. Adding another image induces an increment of 20% with respect to the single-shot case. It is interesting to note that the results for MvsM lie in between these two figures.

In ETHZ dataset, PLS [222] produces the best performances on this dataset. In the single-shot case, the experiments are carried out exactly as for iLIDS. The multiple-shot case is carried out considering $N = 2, 5, 10$ for MvsS and MvsM, with 100 independent trials for each case. Since the images of the same pedestrian

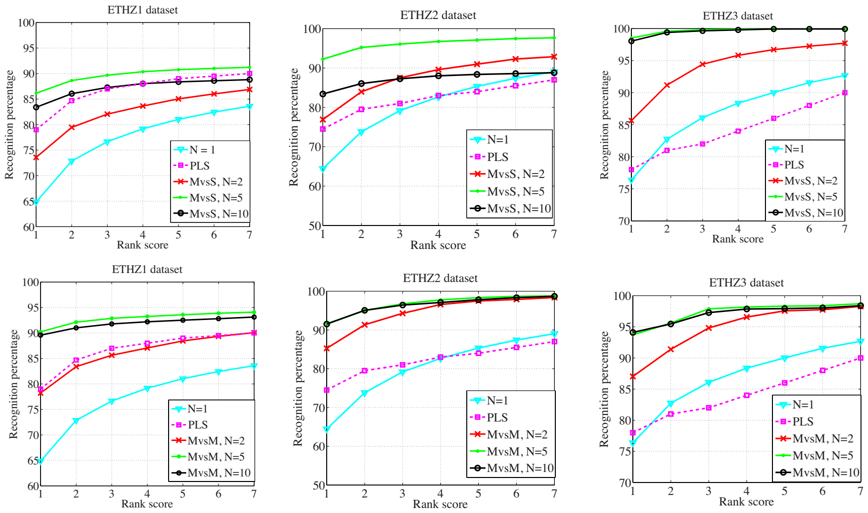


**Fig. 4.13.** Performances ETHZ dataset. Left column, results on Seq. #1; middle column, on Seq. #2; right column, on Seq. #3. We compare our method with the results of PLS [222]. On the top row, we report the results for single-shot SDALF ($N = 1$) and MvsS SDALF; on the bottom row, we report the results for MvsM SDALF. In accordance with [222], only the first 7 ranking positions are displayed.

come from video sequences, many are very similar and picking them for building the multi-shot signature would not provide new useful information about the subject. Therefore, we apply beforehand the clustering procedure discussed in Section 4.3.1.

The results for both single and multiple-shot cases for Seq. #1 are reported on Fig. 4.13, and we compare the results with those reported by [222]. In Seq. #1 we do not obtain the best results in the single-shot case, but adding more information to the signature we can get up to 86% rank 1 correct matches for MvsS and up to 90% for MvsM. We think that the difference with PLS is due to the fact that PLS uses all foreground and background information, while we use only the foreground. Background information helps here because each pedestrian is framed and tracked in the same location, but it is not valid in general in a multi-camera setting. In addition, PLS requires to have all the gallery signatures beforehand in order to estimate the weights on the appearance model. So, if one pedestrian is added the weights must be recomputed and this is another drawback of this technique, weakening its use in real scenarios.

In Seq. #2 (Fig. 4.13) we have a similar behavior: rank 1 correct matches can be obtained in 91% of the cases for MvsS, and in 92% of the cases for MvsM. The results for Seq. #3 show instead that SDALF outperforms PLS even in the single-shot case. The best performances as to rank 1 correct matches is 98% for MvsS and 94% for MvsM. It is interesting to note that there is a point after that adding more information does not enrich the descriptive power of the signature any more. $N = 5$ seems to be the correct number of images to use.

### 4.7.3 (A)HPE

The quantitative evaluation of HPE and AHPE considers the six multi-shot datasets: ETHZ 1, 2, and 3, iLIDS for re-id, iLIDS$_{\geq 4}$, and CAVIAR4REID.

We reproduce the same multi-shot experimental settings described in the previous section for SDALF. We randomly select a subset of $N$ images for each person to build the gallery set, and $N$ for each person for the probe set. Whereas a pedestrian has less than $2N$ images in total (*e.g.*, in iLIDS the individuals with just 2 images are 18.5%), we build the signatures splitting in equal proportions the images for the probe and the gallery. When just 2 images are available, the descriptor becomes single-shot. Then, the matching between (A)HPEs of the probe set and the ones of the gallery set is estimated. To have a robust statistics, this whole procedure is repeated 20 times, and the CMC curves are averaged over the trials.

There are three aspects that we investigate with our experiments: i) we test the HPE descriptor varying the number of images $N$ for each individual, to see how important is to have multiple instances per person. ii) We compare (A)HPE with the state-of-the-art methods for better understanding pros and cons of our proposal. iii) We perform an analysis of the weights $\beta$s in order to find out which feature is more discriminant.

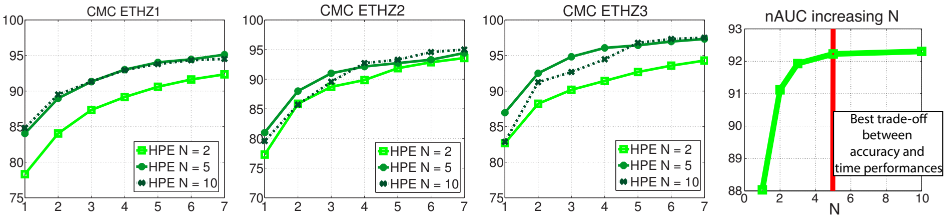


**Fig. 4.14.** Evaluation on ETHZ 1,2,3 of HPE varying the number of images (first three columns). Normalized AUC averaged on ETHZ and iLIDS at increasing the number of images (last column).

First, we analyze HPE varying $N$. We focus on the ETHZ dataset (Fig. 4.14, first three columns) which has enough samples, setting $N = \{2, 5, 10\}$. Due to the nature of the datasets, the results prove that our method is robust to occlusions and quite crowded scenarios (*e.g.*, the images often contain more than a person). Moreover, the analysis of the nAUC (Fig. 4.14, right) shows that the accuracy increases sub-linearly with the number of images $N$. The trade-off between accuracy and time performances is provided by $N = 5$. We could use more images, but the computational time of the matching would increase significantly (it is quadratic in $N$), with a small gain in accuracy. Only ETHZ has such a number of images per person, while for iLIDS we have to set $N = 2$ as in Fig. 4.15. In other words, for iLIDS we cannot exploit our method at its best. In fact, learning epitomes with $N < 5$ is quite tricky because the model over-fits the data and it is not able to generalize a common structure between the views. This effects is even more dramatic with $N = 1$, where performances are very low. In other words, our approach has to be intended solely as multi-shot approach for re-identification.

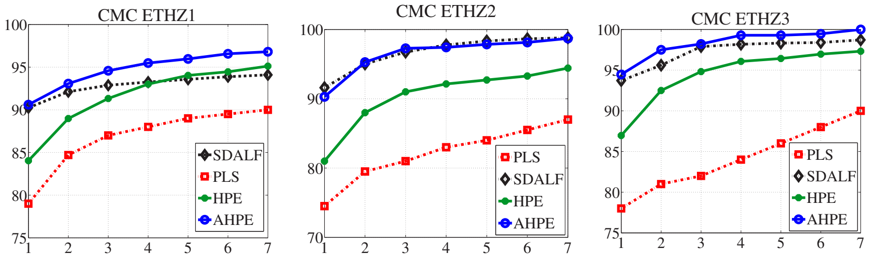


**Fig. 4.15.** Comparisons on ETHZ 1,2,3 between AHPE (blue), HPE (green), SDALF (black), PLS [222] (red). For the multi-shot case we set $N = 5$.

A comparison between different state-of-the-art methods, HPE and AHPE descriptor is shown in Fig. 4.15. On ETHZ, AHPE gives the best results, showing consistent improvements on ETHZ1 and ETHZ3. On ETHZ2, AHPE gives comparable results with SDALF, since the nAUC is 98.93% and 98.95% for AHPE and SDALF, respectively. Note that if we remove the image selection step (used for ETHZ), the performances decreases of 5% in terms of CMC, because the intra-variance between images of the same individual is low, and thus the multi-shot mode does not gain new discriminative information.

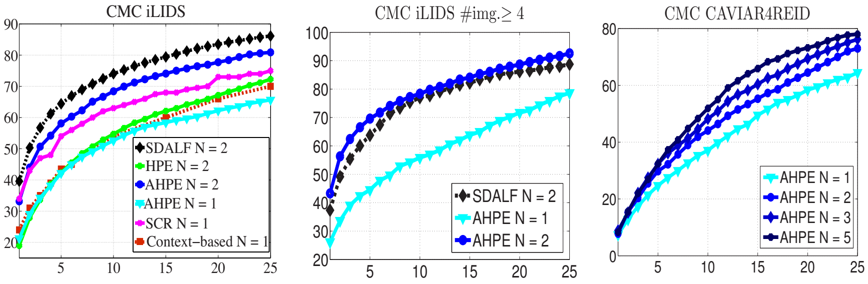
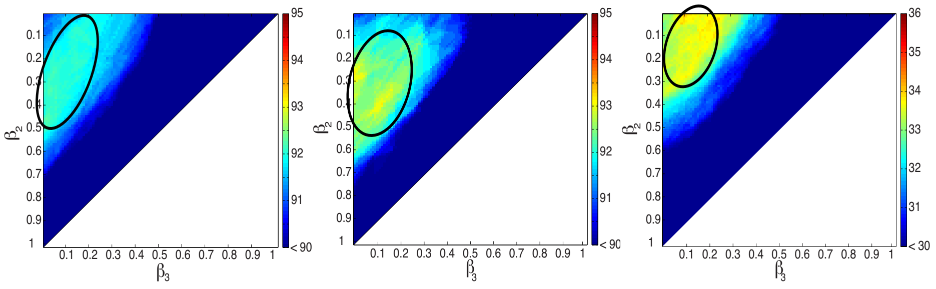


**Fig. 4.16.** Comparisons on iLIDS (first column), iLIDS$_{\geq 4}$ (second column) and CAVIAR4REID (third column) between AHPE (blue), HPE (green, only iLIDS), SDALF (black), SCR [20] (magenta, only iLIDS), and context-based [275] (red, only iLIDS). For iLIDS and iLIDS$_{\geq 4}$ we set $N = 2$. For CAVIAR4REID, we analyze different values for $N$. Best viewed in colors.

On iLIDS (Fig. 4.16, left), AHPE is outperformed only by SDALF. This witnesses again the fact, explained in the previous experiment, that the epitomic analysis works very well when the number of instances is appropriate (say, at least $N = 5$). This statement is clearer by the experiments on iLIDS$_{\geq 4}$ and CAVIAR4REID (Fig. 4.16, last two columns). Especially, if we remove from iLIDS the instances with less than 4 images, then AHPE outperforms SDALF (Fig. 4.16, center). The evaluation on CAVIAR4REID (Fig. 4.16, right) shows that: 1) as HPE in Fig. 4.14 the accuracy increases with $N$, and 2) the real, worst-case scenario of re-identification is still very challenging and an open problem.



ETHZ1 ETHZ2 iLIDS

**Fig. 4.17.** Analysis of the parameters $\beta_{1,2,3}$ in terms of the first rank of the CMC curve (CMC(1)). The black ellipses highlight the optimal parameters for each dataset. The main work is done by the wHSV descriptor, but using the epitome-based features in combination increase the accuracy. Note that values below 90 (for ETHZ) and 30 (for iLIDS) are all dark blue for better visualization and the right-bottom corner is white because the parameters sum to 1.

The last analysis (Fig. 4.17) concerns the evaluation of the performances varying the weight $\beta$s of Eq. 4.12. Note that, we performed a greedy search of the parameters $\beta_i$, but a possible extension of the work could be to learn the weights in a offline/online training.

The quantitative analysis has been performed using the values of CMC at first position (CMC(1)). Fig. 4.17 shows the results for ETHZ1, ETHZ2 and iLIDS varying $\beta_2$ and $\beta_3$ (the value of $\beta_1$ can be derived by $\sum_i \beta_i = 1$). First of all, it is worth noting that if we use just the local epitome or the generic epitome the performances are not the best. Using only the color histogram (the upper-left corner) gives good performances, but again not the best. The best performance are highlighted for each dataset (with ellipses) in Fig. 4.17. This parameters optimization shows that there does not exist a unique set of parameters for all the dataset. Instead, we need to find a trade-off, for example, by intersecting the regions where the accuracy is good. In fact, we can notice that a good choice of the parameters is: $\beta_1 = 0.6$, $\beta_2 = 0.25$ and $\beta_3 = 0.15$. We used this parameters setting in our experiments.

### 4.7.4 Multi-target Tracking with SDALF

In this section, we experiment how the SDALF descriptor performs when dealing with tracking problem. Among the several benchmark datasets available for multi-target tracking, we adopt CAVIAR [1] because it represents a real challenging scenario due to pose, resolution and illumination changes, and also occlusions. In addition, the provided tracking ground truth enables us to perform quantitative

| | HSV | partHSV | SDALF |
|---|---|---|---|
| EnterExitCrossingPaths1cor | 51.74 | 48.61 | **80.90** |
| EnterExitCrossingPaths2cor | 41.49 | 33.20 | **80.29** |
| OneLeaveShop1cor | 60.70 | 49.44 | **67.34** |
| OneLeaveShop2cor | 73.30 | 83.57 | **92.14** |
| OneLeaveShopReenter1cor | 63.50 | 53.85 | **88.72** |
| OneLeaveShopReenter2cor | 47.14 | 50.71 | **65.60** |
| OneShopOneWait1cor | 56.04 | 51.79 | **76.58** |
| OneShopOneWait2cor | 49.61 | 43.83 | **79.20** |
| OneStopEnter1cor | 74.76 | 77.92 | **95.24** |
| OneStopEnter2cor | 72.61 | 70.81 | **84.09** |
| OneStopMoveEnter2cor | 83.66 | 83.08 | **93.61** |
| OneStopMoveNoEnter1cor | 82.67 | 76.31 | **89.82** |
| OneStopMoveNoEnter2cor | 62.07 | 69.50 | **82.05** |
| OneStopNoEnter1cor | 73.14 | 51.84 | **88.71** |
| OneStopNoEnter2cor | 66.92 | 58.29 | **87.93** |
| ShopAssistant1cor | 79.46 | 80.55 | **90.25** |

**Table 4.2.** Tracking success rate of each sequence selected from CAVIAR shopping centre. Three appearance descriptors have been compared (HSV histogram, part-based HSV histogram and SDALF). In all the sequences, the proposed appearance descriptor outperforms the others.

| | FP | MO | FN | MT | TSR | # Est. | # GT |
|---|---|---|---|---|---|---|---|
| SDALF | **7.74** | 0.29 | **16.09** | 1.02 | **83.91** | **148** | 98 |
| partHSV | 26.70 | **0.14** | 38.54 | **0.25** | 61.46 | 251 | 98 |
| HSV | 24.30 | 0.18 | 35.07 | 0.29 | 64.93 | 220 | 98 |

**Table 4.3.** Quantitative comparison between the descriptors: SDALF, part-based HSV histogram and HSV histogram, in terms of False Positives (FP), Multiple Objects (MO), False Negatives (FN), Multiple Trackers (MT), Tracking Success Rate (TSR) and the number of tracks estimated (# Est.) vs. the number of tracks in the ground truth (# GT).

comparisons. We select a representative subset of the sequences (the 16 videos listed in Table 4.2).

The proposed SDALF-based observation model is compared against two classical appearance descriptors for tracking: joint HSV histogram and part-based HSV histogram (partHSV) [113] where each of three body parts (head, torso, legs) are described by a color histogram.

In order to evaluate the method, we use the measurements presented in Section 2.4.1, that are: False Positives (**FP**), Multiple Objects (**MO**), False Negatives (**FN**), Multiple Trackers (**MT**), Tracking Success Rate (**TSR**) and Mean Error (**ME**). In addition, we provide also an evaluation in terms of number of tracks estimated by our method (# **Est.**) vs. number of tracks in the ground truth (# **GT**). It is an estimate of how many tracks are wrongly generated (for example, because weak appearance models cause tracks drifting).

In Table 4.2, the TSR for each sequence is reported for the three descriptors. Our approach achieves the best TSR in every experiment. The accuracy improvement is between 7% (OneLeaveShop1cor) and 47% (EnterExitCrossingPaths2cor). This means that SDALF descriptor is more robust to appearance changes than the others, providing an accurate hand-crafted descriptor for humans.

Moreover, the overall tracking results[7] reported in Table 4.3 highlight the same behavior of the descriptors. Our approach is better in terms of FP, which means that tracking is performed with higher accuracy (*e.g.*, not too large bounding boxes), and in terms of FN, that is, it is less probable to lose targets. The fact that MO and MT are higher using SDALF descriptor is not necessary saying the the proposed descriptor fails. Instead, it is worth notice from the qualitative analysis provided by Fig. 4.18 and 4.19 and by the video in the additional material that these values are higher because our method deals with partial occlusions. HSV and partHSV are not able to deal robustly with this problem, and therefore tracks are lost when it occurs, giving a lower MO and MT but also a lower TSR. In fact, SDALF outperforms HSV and partHSV in terms of overall TSR. The gain in accuracy of our approach is about 19% and 22% with respect to HSV and partHSV, respectively. Analyzing the number of estimates vs. the number of GT tracks, it is easy to see that we generate less tracks than the others (at least 33% less) and it is the closest to the real number of tracks. Thus, HSV and partHSV are less robust than SDALF because of illumination, pose, and resolution changes and partial occlusions. Using these descriptors, several tracks are frequently lost and reinitialized.

A qualitative analysis that highlights the performances discussed above is provided by Fig. 4.18 and 4.19 and the videos reported at `http://www.youtube.com/watch?v=JiW2unf5gwg`. The sequence of Fig. 4.18 (top) shows the problem of single-target tracking when dealing with illumination and resolution changes. HSV (third row) and partHSV histograms (second row) are not able to deal properly with these problems even if the sequence is quite simple (without occlusions, simple background and one target). The result is that the target is lost and then reinitialized several times during tracking: three times for partHSV and two times for HSV. On the other hand, our approach is able to cope with these problems and to track the target for the whole sequence without any track hijacking. In Fig. 4.18 (bottom), tracking becomes more challenging, because the appearance model has to face pose changes and partial occlusions. As in the previous figure, HSV and partHSV lose the track several times, especially because of pose variations and occlusion. SDALF outperforms the others, and is not prone to error when a partial occlusion occurs. A similar behavior is reported in Fig. 4.19. When dealing with pose, illumination, resolution changes and partial occlusions, SDALF outperforms the HSV and partHSV descriptors in terms of less tracker re-initializations and higher accuracy.

In terms of computational speed, we evaluate how long computing Eq. 4.11 takes[8]. Two steps are required: first, the SDALF descriptor for the current hypothesis is extracted, second, the distances on Eq. 4.11 are computed. The first

[7] These values have been computed by averaging the results over all the sequences.

[8] The following values have been computed using our non-optimized MATLAB code on a quad-core Intel Xeon E5440, 2.83 GHz with 4 GB of RAM.

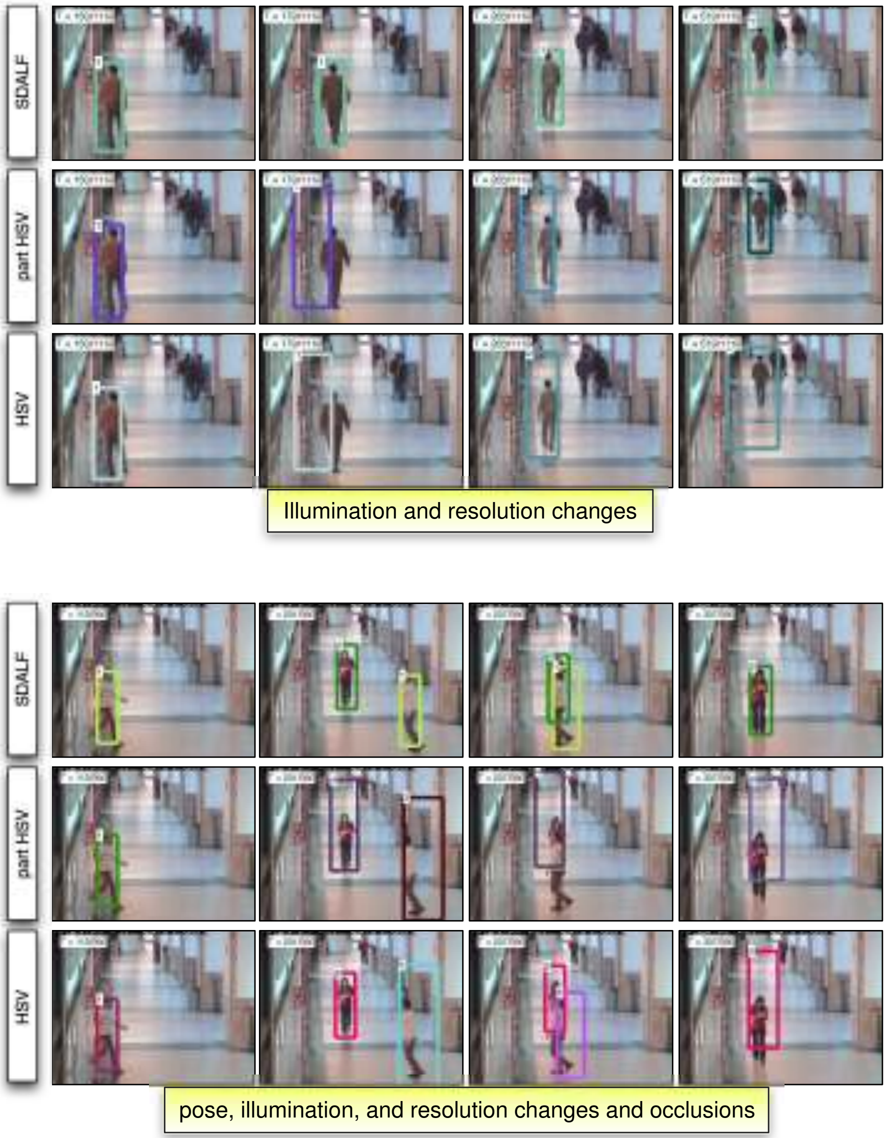

SDALF
part HSV
HSV
Illumination and resolution changes
SDALF
part HSV
HSV
pose, illumination, and resolution changes and occlusions

**Fig. 4.18.** Qualitative comparison between the descriptors on the sequence OneLeaveShop2cor (top) and OneLeaveShopReenter1cor (bottom). This subsequence poses the problem of single-target tracking when dealing with illumination and resolution changes.

and second phase take in average 18 and 15 milliseconds, respectively, when the hypothesis has size 12 × 36. When the hypothesis increases his size to 40 × 46, these phases take in average 26 and 24 milliseconds, respectively. Let $N$ are the

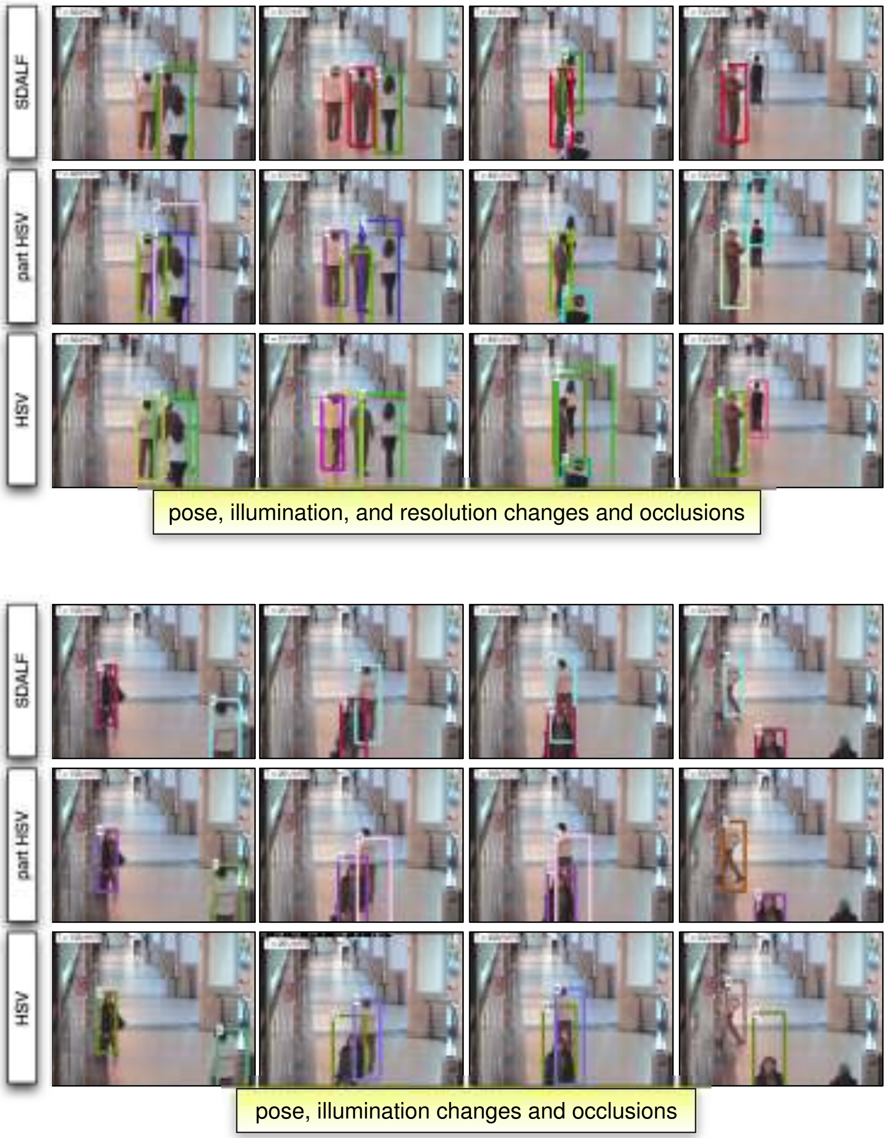

SDALF
part HSV
HSV
pose, illumination, and resolution changes and occlusions
SDALF
part HSV
HSV
pose, illumination changes and occlusions

**Fig. 4.19.** ShopAssistant1cor (top) and OneShopOneWait1cor (bottom) present challenging issues such as pose, illumination and resolution changes and occlusions.

number of particles, $M$ the number of images and $K$ the average number of targets ($M = 3$ and $N = 100$ in our experiments), then the computational complexity of Eq. 4.11 is $\mathcal{O}(K \cdot N \cdot M)$.

## 4.8 Conclusions

In this chapter, several novel methods for both re-identification and tracking have been introduced. First of all, we established a standard pipeline for the re-identification task. In our setting, three descriptors have been proposed: SDALF, HPE and AHPE. We have also proved that SDALF is a robust descriptor that characterizes the human appearance for tracking applications.

SDALF consists in different kinds of features, subsequently combined to generate a robust and discriminative signature, which is then used in a matching strategy for recognition. In particular, perceptual relevant human parts are localized driven by asymmetry/symmetry principles, and three complementary kinds of features are extracted. Each type of feature encodes different information, namely, chromatic and structural information, as well as recurrent high-entropy textural characteristics. In this way, robustness to low resolution, pose, viewpoint and illumination variations is achieved. SDALF resulted to be versatile, being able to work using a single image of a person (single-shot modality), or several frames (multiple-shot modality), in both signature generation and testing phases. We proved the goodness of SDALF descriptor on different challenging public databases for re-identification, and as appearance model of a standard multi-target tracking algorithm. Moreover, SDALF also showed to be robust to very low resolutions, maintaining high performances up to $11 \times 22$ windows size.

As for the tracking issue, we also tested SDALF using a benchmark dataset of videos in comparison with widely employed descriptors. Also in these cases, we achieved better performances in terms of different tracking quality metrics, like the accuracy (tracking success rate), and false positive and false negatives rates.

HPE is based on a collection of global and local features, that embeds information from multiple images per person, showing that the presence of several occurrences of an individual is very informative for re-identification. HPE is composed by 1) a chromatic characterization of the individual appearance (the color histogram), prove to be effective also in SDALF, and 2) the epitome. It collapses a set of images, through a generative model, into a small collage of overlapped patches embedding the essence of the textural, shape and appearance properties of the data. These two ingredients together have shown very good performance in the experiments.

AHPE is a simple extension that merge together the robust description given by HPE with the perceptual principles for silhouette segmentation employed in SDALF. This gives birth to the best descriptor as shown in the comparative analysis.

Overall, a basic question raises for such kind of problems: how can one choose which descriptor is better to use under certain conditions? It depends on the context. Even if AHPE has been identified as the descriptor that gives the best results, each descriptor shows advantages and drawbacks. In the following, we report some guidelines. SDALF is preferred when only few images (say $1-3$) are available for each individual and also in case of very low resolution. (A)HPE has to be chosen when the number of images is sufficiently high (say $\geq 5$). When the images are characterized by low texture or saturated colors (for example, in case of shadows), SDALF performs better because the epitome will not capture much reliable information. In general, a high number of heuristics should be evaluated in order to choose which one is better. Another option could be to run the algorithms in parallel on different computers and leave to the human the final choice.

Analyzing again the results reported in the experimental section, the reader can easily notice that the re-identification algorithms are very far from being actually ready for a real automatic system. The rank 1 of the CMC curves should be near 100% but this is not the case. However, several solutions can be proposed from a practical point of view: 1) Use the re-identification methods to prune out the number of matchings that an human operator has to evaluate; in this way, we use the re-identification methods in order to alleviate the workload and the false negative rate of the human operator. 2) We can even "include the human in the loop": after having run the re-identification method, let the results be evaluated by a human expert, so that to increase the performances of the system. We leave these options as future work that has to be investigated to develop effective re-identification systems.

# Part III

# Analyzing Groups

# 5

# Social Interaction Detection

The modeling of social interaction is becoming a very active trend in the computer vision and social signal processing research in the last decade. Some computer vision examples are: data association [191] and dynamic constrains for tracking [190, 270], person recognition from still images [257], activity recognition [55], and so forth. The basic idea is to use standard computer vision tools (*e.g.*, a detector and a tracker) to extract some high-level, abstract information on the dynamics of the people in the scene. In Figure 5.1, this idea is depicted as an extension of the tracking graphical model: the classical model is connected to an "group" variable $\mathbf{G}_{t+1}$ that models the social entities. The goal is to estimate $\mathbf{G}_{t+1}$ given the tracking results $\mathbf{X}_{t-\Delta+1|t}$ ($\Delta$ is a temporal window). The solutions introduced in this chapter embed notions of social psychology into computer vision algorithms with application to surveillance scenarios. In particular, we deal with two interesting problems: 1) *group interaction discovery*, an instance of social interaction detection where the goal is to find groups of individuals in videos, and 2) the *focus of attention* of individuals and groups, *i.e.*, estimating which portions of a scene that are the most observed. Each proposed method comes with an experimental section on real data, that gives an evaluation of its performance and potentialities.

## 5.1 Introduction

Recently, researchers in surveillance shifted their attention from the monitoring of a single person to that of groups: this novel level of abstraction in surveillance provides event descriptions which are semantically more meaningful, highlighting barely visible relational connections that exist among humans. Even if computer vision and pattern recognition supported this new perspective by providing computational models for capturing the whereabouts of groups, such disciplines rarely consider that the basic ingredient of a group is the human being, and that a group is based on interactions between humans. To the best of our knowledge, all the work that deal with groups assumes that people are simple points on a plane [108, 139, 154, 177, 188, 246, 260] that in some cases may obey to physical laws of attraction and repulsion [190, 223]. None of them considers that working on groups implies to focus on the analysis of the human behavior – a process subject

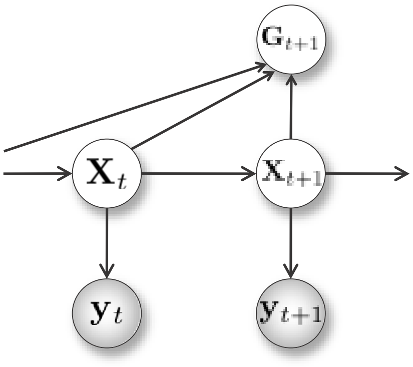


**Fig. 5.1.** Model of group interaction discovery, that gathers the tracking results in a fixed time window to perform inference.

to principles and laws rigorous enough to produce stable and predictable patterns corresponding to social, emotional, and psychological phenomena. On the other hand, these topics are the main subjects of other computing domains, in particular social signal processing and affective computing [251], that typically neglect scenarios relevant to surveillance and monitoring.

Social signal processing and computer vision are tightly intertwined. In our context, they attempt to discover social interactions using statistical analysis of spatial-orientational arrangements that have a sociological relevance. Social signals are conveyed, often outside conscious awareness, by nonverbal behavioral cues like facial expressions, gaze, vocalizations (laughter, fillers, back-channel, etc.), gestures and postures. So, there have been identified a large number of behavioral cues carrying social meaning, which are grouped into five classes called codes [250]: physical appearance (attractiveness, clothes, ornaments, somatotype, etc.) [208,262], vocal behavior (everything else than words in speech) [200,221], face and eyes behavior (expressions, gaze, head pose, etc.) [58,72], gestures and postures (hand and body movements, conscious and unconscious gestures, orientation with respect to others, etc.) [192,220], space and environment (mutual distances, spatial organization of people, territoriality, geometric constraints) [131, 208]. Social interactions are here intended as the acts, actions, or practices of two or more persons mutually oriented versus each other, that is, every behaviors affecting or considering others' subjective experiences or intentions [215]. For instance, talking is the most common kind of social interaction, but working together, playing chess, eating at a table, and offering a cup of water are social interactions too. In general, any dynamic sequence of social actions between individuals (or groups) that modify their actions and reactions by their interaction partner(s) are social interactions.

The methods presented in this chapter take into account these cues in order to give a spectra of algorithms that deal with the *group* entity in a more principled way. In sociology, a group may be defined as a collection composed by a number of individuals who share certain aspects, interact with one another, accept rights and obligations as members of the group and share a common identity. We are conscious that identifying such complex relations is a hard task, if we consider just a video as input. For this reason, we consider to be correct the use of a less constrained meaning of the term group, that is, an assemblage of people standing near together, and forming a collective unity. In particular, we consider the *life of a group*, analyzing how the presence of a group can be detected in crowded situations (i.e., the birth and the death of a group), and what are the basic activities carried out by their components in terms of interactions between the humans and the environment. In particular, we analyze what are the zones of the environment where the attention of humans is more focused. In the next chapter, we also analyze how a moving group can be tracked, that is, its *evolution* over time.

The *birth of a group* or its detection (and, consequently, its break-up) can be performed in two ways. The first assumes that individual are stationary for a period of time in a given location: for example, in a cocktail party, people may be discussing for a while around a table, before leaving. In a canteen, elements of a group may be clustered around a vending machine. In these cases, social theories help in individuating a group, which can be subsequently followed by a tracking algorithm. In particular, relative positioning and head direction may support this analysis. The second method takes into account situations where people usually move, and no aggregations of stationary people may be observed. In this case, we advocate the use of proxemics, which states that the kind of relation present among the persons depends on the distances they have with respect to each other. Here, we mainly take into account the first scenario, where more interesting and stable (over time) social interactions can be found.

Moreover, we present the idea of how people interact with an environment through an *interest map*, *i.e.*, a map that highlights the parts of a scene more attended by a person. For instance, a vending machine in an empty room will surely attract more the attention of people than the peripheral walls. In this direction, we present a system that exploits the use of the head position and orientation for extracting such information.

All these aspects that characterize a group build upon unconventional features that change the perspective followed so far by researchers involved in video surveillance: from the general and unique point of view of a single camera mounted on a wall to a *subjective*, personal, viewpoint, aimed at understanding what is experienced by each single person in the monitored scene. In this context, we propose a general social scenario in which we estimate the position of every person so as to keep track of the related distance among them. This will help in inferring the kind of relation which holds among people in a scene. Another interesting which is exploited is the visual focus of attention, that is the visual field of view of a person approximated using computational geometry techniques. This helps in estimating the focus of attention of a person while immersed in a whatever scenario.

The rest of the chapter is organized as follows. Section 5.2 details the background computer vision tools used in the proposed method. The core of the chapter is represented by Sections 5.3 and 5.4, describing an approach to initialize the groups, and a way to create interest maps of a monitored setting, respectively. Qualitative and quantitative analysis is provided in Section 5.5. Finally, a discussion follows in Section 5.6. On the other hand, for the analysis of the group evolution over time, the reader should wait for the next chapter.

## 5.2 Background

The automatic recognition of social interactions in video recordings is undoubtedly one of the main challenges for a surveillance system. This is usually accomplished using a serial architecture built upon an array of techniques aimed at extracting low-level information, followed by a classification stage. Computer vision techniques are typically exploited for extracting these low-level features from videos, useful to allow high-level inference. First, all the people in the camera-monitored environment have to be localized, and tracking algorithms are employed to provide the position of each person at each time instant, *i.e.*, its trajectory. When the position is estimated, a head orientation method computes the pose of the head of each person, and the subjective field of view (also called the view frustum) is localized by exploiting the calibration of the camera. Thus, the basic components used by the proposed architecture are: a multi-person tracker (HJS filter described in Section 2.2.1), a head pose estimation method (Section 5.2.1), and the subjective view frustum estimation (Section 5.2.2). We discuss below these last two components of the system and we refer to Section 2.2.1 for details about the multi-person tracker.

### 5.2.1 Head Orientation Estimation

Several head orientation estimation approaches have been proposed in the literature: a recent review can be found in [173], where a performance analysis of different methods is presented, and a list of the commonly used dataset for head pose estimation is shown. Moreover, the CLEAR workshops are important events for the head pose estimation community, and several important approaches can be found in the related proceedings [233, 234]. It is worth noting that most of the approaches are based on classification schemes. In the multi-faceted ensemble of the classification approaches, boosting-based techniques play a primary role [149, 182, 243, 252, 266, 268]. Boosting [89, 219] is a learning algorithm that creates strong and fast classifiers, employing various features fed into diverse architectures with *ad-hoc* policies. Among the different features exploited for boosting in surveillance applications (see [267] for an updated list), covariance features [242] have been exploited as powerful descriptors of pedestrians [243, 266], and their effectiveness has been investigated in [182]. When injected in boosting systems [182, 241, 243, 266], covariances provide strong detection performance, encapsulating possible high intra-class variances (due to pose and view changes of an object of interest). They are in general stable under noise, and furnish an elegant way to fuse multiple low-level features as, in fact, they intrinsically exploit possible inter-feature dependencies. For these reasons, in this work we are going to use boosting combined with the covariance matrix descriptor.

The tracker provides the feet position in the ground floor for each person in each frame. Using the calibration parameters, a cylindric model built on that position is back-projected to the image plane. This gives the approximate position of the head. We define a square window $I$ of size $r \times r$, where we run a multi-class algorithm that recovers the head orientation. The size $r$ is chosen large enough in order to contain a head, considering the experimental physical environment and the camera position. Note that another possible solution would be to directly track the head of the persons and project it to the ground floor in order to estimate the person position.

As multi-class classification, we combine boosting and regression trees [42,241], because they are the ideal weak learning strategy, since they can tolerate a significant amount of labeling noise and errors in the training data (which are very likely in low resolution images). Moreover, they are very efficient at runtime, since matching a sample against a tree is logarithmic in the number of leaves.

From the mathematical point of view, they are an alternative approach to nonlinear regression. The principle is to subdivide, or partition, the space in two smaller regions, where the data distribution is more manageable. This partitioning proceeds recursively, as in hierarchical clustering, until the space is so tight that a simple model can be easily fitted. The global model thus has two parts: one is just the recursive partition, the other is a simple model for each cell of the partition. Regression trees are more powerful than global models, like linear or polynomial regression, where a single predictive formula is supposed to hold over the entire data space. In order to avoid the risk of overtraining of the regression tree, we establish as stopping rule a minimal number $\tau$ of observations per tree leaf, that is experimentally estimated (Section 5.5).

In our approach, we extract from each image of size $I$ ($r \times r$ pixels), a set of dimension $d = 12$ features $\Phi(I, x, y)$ where $x, y$ are the pixel locations, that is defined as follows:

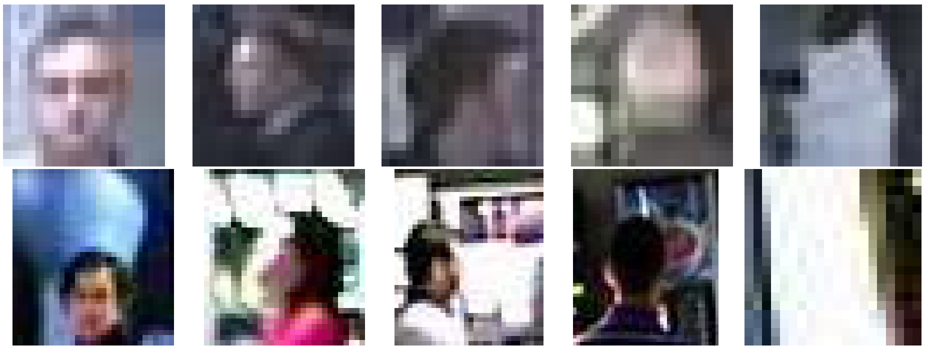
$$\Phi(I,x,y) = \left[X \; Y \; R \; G \; B \; I_x \; I_y \; O \; \mathrm{Gab}_{\{0,\pi/3,\pi/6,4\pi/3\}}\right] . \tag{5.1}$$

$X, Y$ represent the spatial layout maps in $I$, and $R, G, B$ are the color values in the RGB space. $I_x$ and $I_y$ are the directional derivatives of $I$, and $O$ is the gradient orientation. Finally, Gab represents the Gabor filters: we use a set of 4 maps containing the results of the filtering, with filters of dimension $2 \times 4$, sinusoidal frequency 16, and directions $\mathcal{D} = \{0, \pi/3, \pi/6, 4\pi/3\}$. In order to increase the robustness to local illumination variations, we apply the normalization operator introduced in [243] before applying the multi-class framework. First, we estimate the covariance of the image $I$, denoted as $X_I$. Then, for each element $X_i$ of the dataset, we apply the following normalization:

$$\widehat{X}_i = \mathrm{diag}(X_I)^{-\frac{1}{2}} X_i \, \mathrm{diag}(X_I)^{-\frac{1}{2}}, \tag{5.2}$$

where $\widehat{X}_i$ is the normalized descriptor, and $\mathrm{diag}(X_I)$ is a square matrix with only the diagonal entries of $X_I$.

Our approach takes inspiration from the literature on dense image descriptors [65]. We sample the window $I$ employing an array of uniformly distributed and overlapping patches of the same dimension. For each of the $N_P = 16$ sampled patches inside the $r \times r$ region of interest, described by the covariance matrix of a set of $d$ image features described by the Eq. 5.1, a multi-class LogitBoost classifier is trained. Each class represent a different head orientation sampled according with a fixed sampling step $\alpha$ and from an extra class containing all the background examples. We experimentally found that $\alpha = 90°$ which correspond to the semantic classes North, South, East and West, is enough for our purposes. We also add



(a) North (b) East (c) West (d) South (e) Background

**Fig. 5.2.** Examples of the 5 semantic classes we defined for the multi-class problem of head pose estimation. a) North, b) East, c) West, d) South, and e) Background. The first row shows some examples of the training set, and the second row shows some sample windows at testing time. Note that the images have very low resolution (min. 20 pixels).

the Background class to minimize the false positive rate when the tracker fails to provide a correct head position. We are aware that the use of only four directions may lead to rough estimates, but it should be considered that the resolution of the source video data is very poor. Fig. 5.2 shows some training and testing examples for each class. At testing time, each patch of a sample window (Fig. 5.3) is independently classified. Then, the classification result is given by a majority criterion across the patches. We name the combination of this patch description that encodes the local shape and appearance and its uniformly distributed architecture *ARray of COvariances* (ARCO).

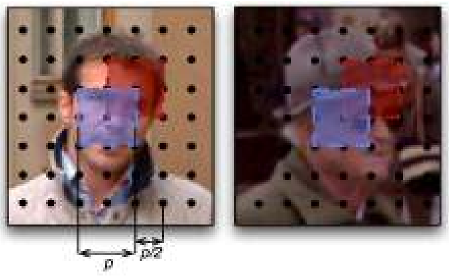


**Fig. 5.3.** Array of Covariance matrices (ARCO) feature. The image is organized as a grid of uniformly spaced and overlapping patches. The head orientation result of each patch is estimated by a multi-class classifier.

More formally, given a set of patches $\{P_i\}_{i=1,\ldots,N_P}$, we learn a multi-class classifier for each patch location $\{F_{P_i}\}_{i=1,\ldots,N_P}$ through the multi-class LogitBoost algorithm [89], adapted to work on Riemannian manifolds, as suggested by [241, 243]. This method implies that each covariance matrix must be projected on a proper tangent space (vector space) of the Riemannian manifold to be classified. Since we deal with a multi-class problem, a common tangent space is chosen where all the covariances are projected and discriminated. Computational considerations suggest to use the identity matrix $I_d$ as projection point. From a mathematical point of view, the projection is a logarithmic transformation of the (positive) eigenvalues of a covariance matrix; therefore, the computational complexity of each projection is bounded by the eigenvalue decomposition complexity $O(d^3)$. Since $d$, the number of image features, is small the projection results a fast operation. All the details of the projection operation are contained in [241, 243].

Let $\Delta_j = \sum_{i=1}^{N_P}(F_{P_i} == j)$ be the number of patches that vote for the class $j \in \{1, \ldots, J\}$. To assign a class label $c$ to a new image, we fuse the votes with a majority voting strategy among all the classes:

$$c = \arg\max_j\{\Delta_j\}, \quad j = 1, \ldots, J. \tag{5.3}$$

The ARCO representation has several advantages. First, it allows to take into account different features, inheriting their expressivity and exploiting, by definition, possible correlations. Second, due to the use of integral images for the computation of the covariance matrices [243], ARCO is fast, making it suitable for a possible real-time usage.

### 5.2.2 Subjective View Frustum Estimation

The Visual Focus Of Attention (VFOA) [159,227,236] is a very important aspect of non-verbal communication. It is well known that a person's VFOA is determined by his eye gaze. Since objects are foveated for visual acuity, gaze direction generally provides more precise information than other bodily cues regarding the spatial localization of the attentional focus. A detailed overview of gaze-based VFOA detection in meeting scenarios is presented in [16]. However, measuring the VFOA by using eye gaze is often difficult or impossible: either the movement of the subject is constrained or high-resolution images of the eyes are required, which may not be practical [165, 229], and several approximations are considered in many cases. For example, in [236], it is claimed that the VFOA can be reasonably inferred by head pose, and this is the choice made in many works. Following the same hypothesis, in [227] pan and tilt parameters of the head are estimated, and the VFOA is represented as a vector normal to the person's face, and it is employed to infer whether a walking person is focused on an advertisement located on a vertical glass or not. Since the situation is very constrained, this proposed VFOA model works pretty well, but a more complex model, considering camera position, person's position and scene structure, is required in more general situations. The same considerations hold for the work presented in [159], where Active Appearance Models are fitted on the face of the person in order to discover which portion of a mall-shelf is observed. In [137], the visual field is modeled as a tetrahedron associated with a head pose detector. However, their model fixes the depth of the visual field, and this is quite unrealistic.

In cases where the scale of the scene does not allow to capture the eye gaze directly, viewing direction can be reasonably approximated by just measuring the

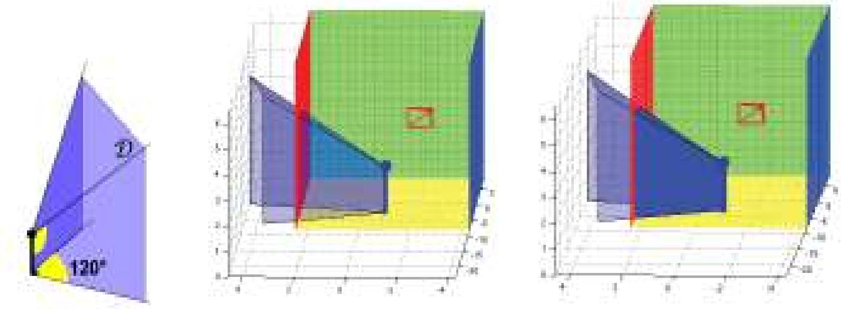


**Fig. 5.4.** Left: the SVF model. Center: an example of SVF inside a 3D "box" scene. In red, the surveillance camera position: the SVF orientation is estimated with respect to the principal axes of the camera. Right: the same SVF delimited by the scene constraints (in solid blue).

head pose. This assumption has been exploited in several approaches dealing with a meeting scenario [235, 236, 254] or in a smart environment [141, 227]. Following this claim, and considering a general, unrestricted scenario, where people can enter, leave, and move freely, we approximate VFOA as the *Subjective View Frustum* (SVF), first proposed in [76]. This feature represents the three-dimensional visual field of a human subject in a scene. According to biological evidence [185], the SVF can be modeled as a 3D polyhedron delimiting the portion of the scene that the subject is looking at (Figure 5.4).

More in detail, the SVF is defined as the polyhedron $\mathcal{D}$ depicted in Figure 5.4 on the left. It is composed by three planes that delimit the view angles on the left, right and top sides, in such a way that the angle span is 120° in both directions. The 3D coordinates of the points corresponding to the head and feet of a subject are obtained from a multi-target tracker (the HJS filter in our case), while the SVF orientation is obtained by the head pose detector described in the previous section.

The SVF $\mathcal{D}$ is computed precisely using computational geometry techniques. It can be written as the intersection of three negative half-spaces defined by their supporting planes of the left, right and top sides of the subject. In principle, the SVF is not bounded in depth, modeling the human capability of focusing possibly on a remote point located at infinite distance. However, in practice, the SVF is limited by the planes that set up the scene, according to the 3D scene (Figure 5.4 on the right). The scene volume is similarly modeled as intersection of negative half-spaces consequently, the exact SVF inside the scene can be computed solving a simple *vertex enumeration* problem, for which very efficient algorithms exist in literature [198].

## 5.3 Human-human Interactions: Group Detection

Employing the SVF in conjunction with cues of the *space and environment* category allows to detect signals of the possible people's interest, with respect to both the physical environment [76], and the other participants acting in the scene. More specifically, we present a method to statistically infer if a participant is involved in an interactional exchange. In accordance with social signal processing studies, we define the birth of a group when multiple and stable relations are detected consistently over time. In particular, it is highly probable that a relation takes place when two persons are closer than 2 meters [250], and looking at each other [115, 140, 264]. We assume that this condition can be reliably inferred by the position and orientation of the SVFs of the people involved. This information can then be gathered in a matrix that we called the *Inter-Relation Pattern Matrix* (IRPM). The IRPM encodes the person-to-person social exchanges occurred considering the individuals in a scene.

Detecting human relations may be useful to instantiate a more robust definition of group in surveillance applications. Actually, in the last few years, several applications focused on group modeling [163, 169], and person re-identification [275] have been proposed. In the former case, a group is defined following physically-driven proximity principles. While in the latter, groups are exploited to improve person re-identification, relying on the fact the people usually stay in the same group when moving in an environment.

Our proposal is a step towards automatic inference and analysis of social interactions in general, unconstrained conditions: it is alternative to the paradigm of wearable computing [56, 193], or smart rooms [255]. In the typical non-cooperative video surveillance context or when a huge amount of data is required, wearable devices are not usable. Moreover, the use of non-invasive technology makes people more prone to act normally.

Considering the literature, the "subjective" point of view for automated surveillance systems has been investigated in [37], taking inspiration from [210], and it represents therefore the most similar approach in the literature to our work. The differences between [37] and our system are that 1) in [37], the gaze is projected on the ground plane, while in our case we embed the 3D subjective view frustum in the 3D scene, employing computational geometry rules, so that the full 3D information allows finer spatial reasoning, needed, for example, to deal with head poses having different tilt angles. 2) They do not perform interaction analysis, and the subjective point of view was functional solely on the estimation of scene interest maps.

Summarizing, we introduce the concept of Inter-Relation Pattern Matrix that exploits the SVF. Its aim is to infer relations among people for detecting groups in a general crowded scenario. This work not only fills a gap in the state of the art of social signal processing aimed at understanding social interactions, but also represents a novel research opportunity, alternative to the scenarios considered so far in socially-aware technologies, where automatic analysis techniques for the spatial organization of social encounters are taken into account. In this section, we consider a scenario where individuals are quasi-stationary for a short period of time in a given location, and we use just simple proxemics cues [250], when dealing with moving people. However, the SVF can also be exploited as a supplementary hint to make more robust the proxemics-based method even in that scenario.

In Section 5.3.1, the method to build the Inter-Relation Pattern Matrix is described. Then, in Section 5.5, experiments and results on home-made and public datasets are shown.

### 5.3.1 The Inter-Relation Pattern Matrix

The SVF can be employed as a tool to discover the visual dynamics of the interactions between two or more people. Such analysis relies on few assumptions with respect to social cues, i.e., that the entities involved in the *social interaction* stand closer than 2 meters (covering thus the *socio-consultive zone* – between 1 and 2 meters – the *casual-personal zone* – between 0.5 and 1.2 meters – and the *intimate zone* – around 0.4-0.5 meters) [250]. Then, it is generally well-accepted that initiators of conversations often wait for visual cues of attention, in particular, the establishment of eye contact, before launching into their conversation during unplanned face-to-face encounters [115, 140, 264]. In this sense, SVF may be employed in order to infer whether an eye contact occurs among close subjects or not. This happens with high probability when the following conditions are satisfied: 1) the subjects are closer than 2 meters; 2) their SVFs overlap, and 3) their heads are positioned inside the reciprocal SVFs. An example of positive result of the condition is shown in Figure 5.5. The *IRPM* records when a possible social interaction occurs, and it can be formalized as a three-dimensional matrix [86], where each entry $IRPM(i, j, t) = IRPM(j, i, t)$ is set to one if subjects $i$ and $j$ satisfy the three conditions above, during the $t$-th time instant.

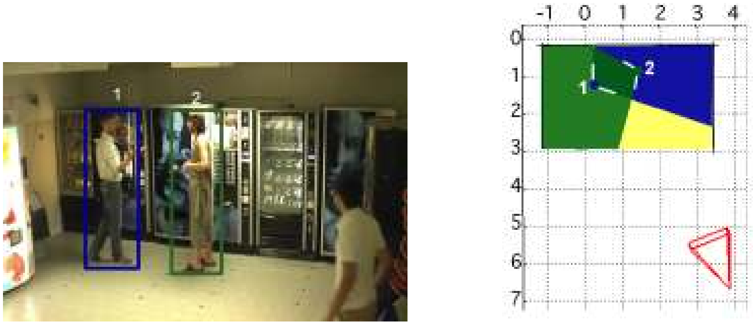


**Fig. 5.5.** Left: two people are talking each other. Right: top view of their SVFs: the estimated orientation, East for 1 and West for 2, is relative to the camera orientation (the pyramid in red in the picture). The SVFs satisfy the three conditions explained in Section 5.3.1.

The IRPM matrix serves to analyze time intervals in which we look for social interactions. Let us suppose to focus on the time interval $[t - T + 1, t]$. In this case we take into account all the IRPM slices that fall in $[t - T + 1, t]$, summing them along the $t$ direction, and obtaining the *condensed* IRPM (*cIRPM*). Intuitively, the higher is the entry $cIRPM_t(i,j)$, the stronger is the probability that subjects $i$ and $j$ are interacting during the interval $[t-T+1,t]$. Actually, *cIRPM* can also be view as a bi-dimensional probability distribution, if normalizes (divide it by $T$). Therefore, in order to detect a relation between a pair of individuals $i,j$ in the interval $[t-T+1,t]$, we check if $cIRPM_t(i,j) > Th$, where $Th$ is a pre-defined threshold. This threshold filters out noisy group detection: actually, due to the errors in the tracking and in the head pose estimation, the lower the threshold, the higher the possibility of false positives detection. In the experiments, we show how the choice of the parameters $T$ and $Th$ impacts on the results, in term of social interaction detection rates.

The cIRPM represents person-to-person exchanges only, but we would like also to capture the presence of *groups* in the scene. Here, we will not use the term group in its sociological meaning, because we are aware that detecting such complex relations using just a video as input is a hard task. For this reasons we consider the group, as an assemblage of people standing near together, and forming a collective unity, a knot of people. The latter meaning is closer to our aims.

Operationally, we treat the *cIRPM* as the adjacency matrix of an undirected graph, with a vertex $v_i$ for each people in the scene, and an edge $e_{ij}$ if $cIRPM_t(i,j) > Th$. The *groups* present in the scene are detected by computing the connected components of the graph.

Using the connected components to determine group is making the assumption that social interaction is a transitive relation. This is not always true in general, where for example the persons at the head and tail of a chain have no interaction at all with each other. A stricker definition of a group wuold require all pairs to be interaction, thus we would be searching for cliques in the graph. However, a video-surveillance system is often prone to error especially when dealing with occlusions. In that case, we will frequently have missing links in the graph, and thus a full clique will occur very rarely. For that reason, we preferred the connected components. Some illustrative examples are depicted in Figures 5.9, 5.10 and 5.11.

## 5.4 Human-Environment Interaction: Interest Map

The contribution of this Section is a visualization application of the SFV-based framework proposed in Section 5.2.2, called the *Interest Map*. Since the part of a scene that intersects the SVF is the area observed by a person, we collect this information over time to infer which are the parts of the scene that are more observed, and thus, where human attention is more plausibly focused. The gathered information is visualized as a suitable color map, in which “hot” colors represent the areas more frequently observed, and the opposite for the “cold” areas. This kind of inference is highly informative at least for two reasons. The first one is diagnostics, in the sense that it gives us the possibility to observe which are the areas of a scene that arouse more attention by the people. The other one is prognostics, since it enables us to devise the parts of the scene that are naturally more observed, because for example they are the natural front of view in a narrow transit area, or for other reasons that this method cannot guess (the interest map only highlights the tangible effects). This application could then be employed for a posterior analysis. In a museum, for example, one may be interested in understanding which artworks receive more attention, or in a market which areas attract more the customers. In a prognostic sense, it may be useful for marketing purposes, such as for example decide where to hang an advertisement.

Section 5.4.1 describes how a 3D map of the monitored environment is created. Since an accurate head pose estimation is not always possible, for example, because of low resolution, an alternative way to describe the pose is the motion orientation of a person (described in Section 5.4.2). The interest map generation process is presented in Section 5.4.3, and experiments, reported in Section 5.5.2, show qualitative results of the interest map given the monitored environment.

### 5.4.1 3D Map Estimation

Let us suppose that the camera monitoring the area is fully calibrated, *i.e.*, both internal parameters and camera position and orientation are known. For convenience, the world reference system is fixed on the ground floor, with the $z$-axis pointing upwards. This permits to obtain the 3D coordinates of a point in the image if the elevation from the ground floor is known. In fact, if $\mathsf{P}$ is the camera projection matrix and $\mathbf{M} = (M_x, M_y, M_z)$ the coordinates of a 3D point, the projection of $\mathbf{M}$ through $\mathsf{P}$ is given by two equations:

$$u = \frac{\mathbf{p}_1^\top \mathbf{M}}{\mathbf{p}_3^\top \mathbf{M}}, \quad v = \frac{\mathbf{p}_2^\top \mathbf{M}}{\mathbf{p}_3^\top \mathbf{M}}, \quad \text{with } \mathsf{P} = \begin{bmatrix} \mathbf{p}_1^\top \\ \mathbf{p}_2^\top \\ \mathbf{p}_3^\top \end{bmatrix}. \tag{5.4}$$

$(u, v)$ are the coordinates of the image point. Thus, knowing $(u, v)$ and $M_z$ it is possible to estimate the position of $\mathbf{M}$ in the 3D space.

Therefore, a rough reconstruction of the area, made up of the principal planes present in the scene, can be carried out (see an example in Figure 5.6). These planes represent the areas of the scene that are interesting to analyze, and the Interest Map will possibly be estimated on them only. Nevertheless, in principle, a more detailed 3D map can be considered, which can be obtained in two ways: first, a manual modeling of the scenario through Computer-Aided Design technologies, and, second and more interestingly, using Structure-from-Motion algorithms [47, 96, 231].

### 5.4.2 Motion Orientation Estimation

The tracking algorithm of Section 2.2.1 provides the position of each person $i$ present in the scene at a certain moment $t$. When it is not possible to apply an head pose estimation algorithm, a simpler pose estimation method is required. In this case, the motion vector can provide the orientation $\theta_{i,t}$ where people are watching. This is a reasonable assumption in a dynamic scenario, because when people walk, they usually look at the direction where they are suppose to go, and therefore they tend to keep the head lined up with the body most of the time. We calculate the angle between the motion direction, given by the tracker, and the camera orientation, using the camera calibration parameters. Therefore,

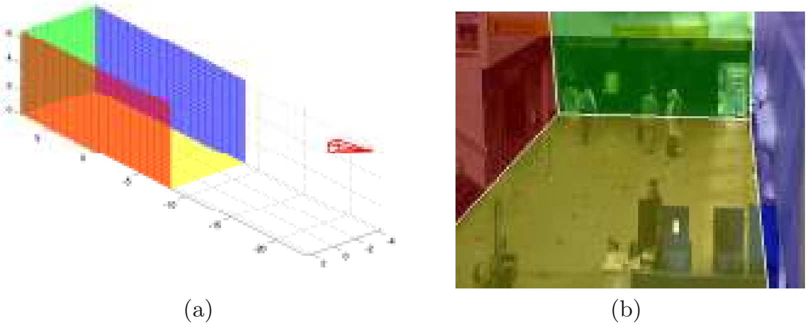


**Fig. 5.6.** 3D reconstruction of the area being monitored. (a) The 3D map of the principal planes. The red cone represents the camera. (b) The planes are projected through the camera and superimposed on one image.

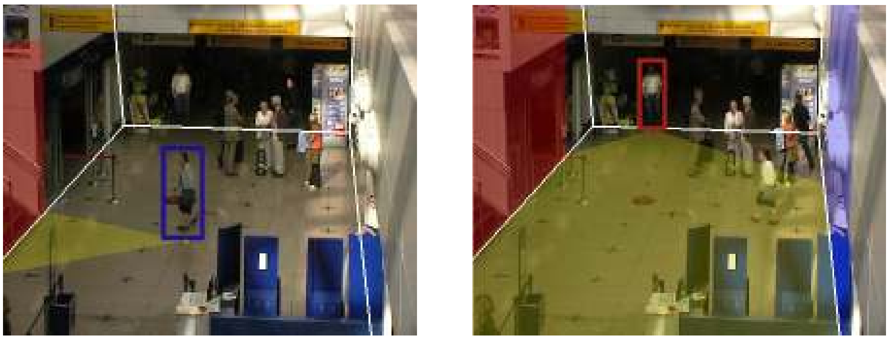


**Fig. 5.7.** Two examples of projection of the SVF on the scene's main planes. The 3D map permits to suitably model the interactions of the SVF with the scene.

this approach can be seen as an alternative, yet simpler solution to the method proposed in Section 5.2.1, that could be useful in specific cases. Moreover, the two approaches could be fused in order to rule out the disadvantages and for making the pose estimation more robust when dealing with both static and moving people.

### 5.4.3 Interest Map Generation

Once we have estimated the ground floor position and orientation of each individual $(x_{i,t}, y_{i,t}, \theta_{i,t})$, we instantiate a SVF for each person. The SVF $\mathcal{D}_{i,t}$ represents the portion of 3D space seen by the $i$-th subject and it is constrained to the main planes of the scene described in Section 5.4.1. A full volumetric reasoning could be considered too, but this would capture other kinds of information, such as people interactions.

Each SVF $\mathcal{D}_{i,t}$ at current time is projected on each scene plane. This is equivalent to estimate the vertices of $\mathcal{D}_{i,t}$ lying on each plane, project these vertices onto the image and select those pixels that lie inside the convex hull of the projected vertices. In this way, the selected pixels represent the projection of each SVF in

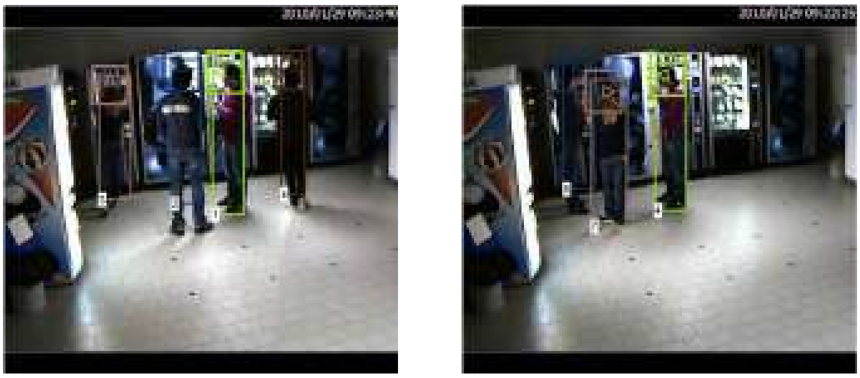


**Fig. 5.8.** Examples of tracking and head orientation classification results. The largest box represents the tracking estimation, the smaller box the area where the head is positioned, and the triangle depicts the estimated head orientation.

the image plane. Two examples of the projected SVF are shown in Figure 5.7. The projections of the SVFs of all the subjects present at the current time-step are then accumulated in a instantaneous map $M_t$ (2D matrix of the same size of the camera frames). We define the *interest map* as the accumulation over time of these instantaneous maps, *i.e.*, IM= $\sum_{t=1}^{T} M_t$. Note that the interest map IM can be computed also in a time window (sum from $T - \tau + 1$ to $T$, where $\tau$ is the size of time window) when the sequences are very long, like in real scenarios. The contributions provided by all tracked people in the sequence, or a set of sequences, are conveyed in the same interest map. Using a similar procedure, a subjective interest map (one independent map for each subject) could easily be computed, but here we restrict the analysis to the interest map for all the subjects. Note that the values of the interest maps vary in the range $[0, K \cdot \tau]$ where K is the number of total tracks and $\tau$ is the chosen size of the time window.

## 5.5 Experiments

In this section, we quantitatively and qualitatively evaluate the methods for group detection and interest map building proposed in Section 5.3 and 5.4, respectively.

### 5.5.1 Group Detection

In order to show the capabilities of the condensed IRPM method, we build our own dataset mainly because most of the public datasets do not contain the ground truth of the social groups in the scene. We recorded a video sequence of about 3 hours and a half duration, portraying a vending machines area where students take coffee and discuss (Figure 5.8). The video footage was acquired with a monocular IP camera, located on an upper angle of the room. The people involved in the experiments were not aware of the aim of the experiments, and behaved naturally.

Afterwards, since creating the ground truth by using only the video is a complex task, we asked to some of them to fill a questionnaire inquiring if they talked and interacted to someone in the room and to whom. Then, the video was analyzed by a psychologist able to detect the presence of interactions among people. The questionnaires were used as supplementary material to confirm the validity of the generated ground truth. This offers us a more trustworthy set of ground truth data for our experiments.

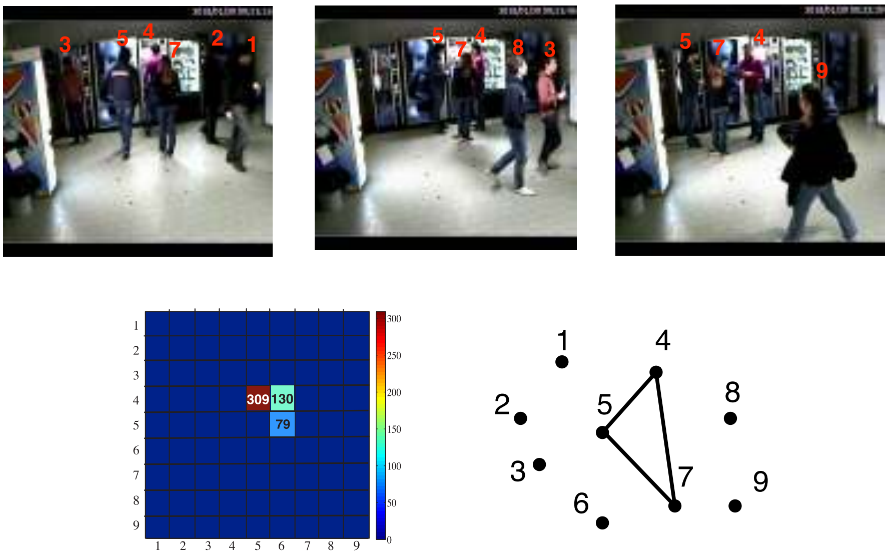


**Fig. 5.9.** Example of IRPM analysis of sequence $S_{04}$. On the top row, some frames of the sequence. On the bottom row, on the left, the cIRPM matrix. Being the cIRPMs symmetric and having null main diagonals, we report for clarity only its strictly upper triangular part. On the right, the corresponding graph. As one can notice, only one group (composed by people 4, 5 and 7) is detected. This is correct, since the other persons in the sequence were not interacting.

The publicly available dataset, called GDet [3], is composed of 12 sub-sequences of about 2 minutes each. They are chosen such that to represent different situations, with people talking in groups and other people not interacting with anyone. For each sub-sequence, we performed tracking, head orientation classification (as shown in Figure 5.8), and built of the three-dimensional IRPM, indicating which people are potentially interacting at a specific moment.

The comparison of our results with the ground truth revealed that 8 out of 12 sequences where correctly interpreted by our system. One can be considered wrong, because there are 2 groups in the scene, and our system reveals that they belong all to the same group. In the other three sequences there are some inaccuracies, like a person left out of a group. These inaccuracies are mainly due to error propagation from tracking and head orientation classification, particularly challenging when people are grouped together and frequently intersect, that is,

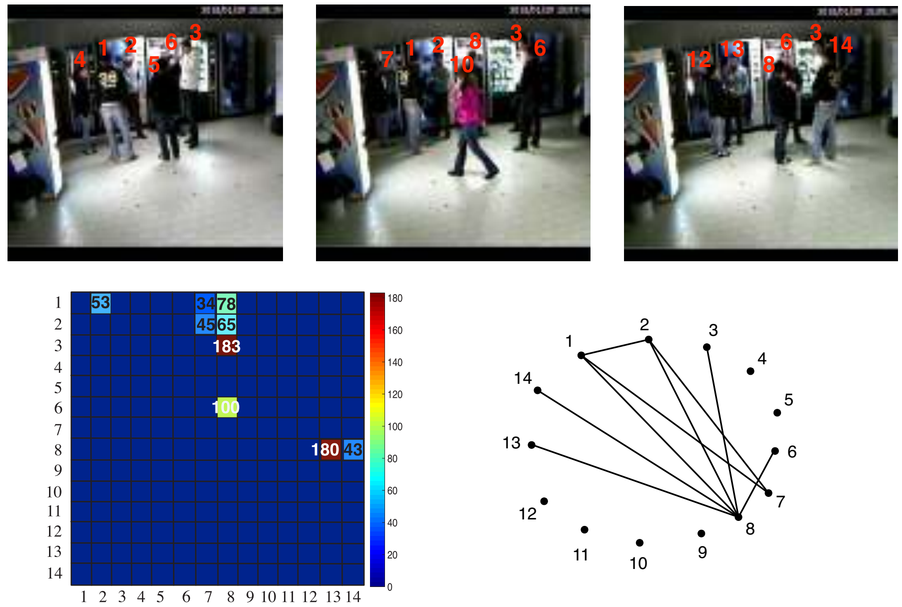


**Fig. 5.10.** Example of cIRPM analysis of sequence $S_{08}$. One big group (1,2,3,6,7,8,13,14) is detected. Note that some people are represented by more than one track, since due to severe or complete occlusions the tracks are sometimes lost and need to be reinitialized (see the text for more details). Person 10 that enters in the room is correctly detected as non-interacting by the cIRPM.

there are many strong, complete occlusions. A qualitative analysis of the results is shown in Figures 5.9, 5.10 and 5.11. The first row of each figure depicts three sampled frames from each sequence and contains the identifiers of each person. The second row depicts the *cIRPM* on the left and the graph structure that defines the group interactions on the right. In all the three experiments, all the groups are detected correctly. In particular, Fig. 5.9 shows the case where a single, small group and other individuals are present in the scene during the recording. In Fig. 5.10, a more complex situation is analyzed, that is, a big group is in the scene (composed by 6 individuals). One big group $(1, 2, 3, 6, 7, 8, 13, 14)$ is found by our method. Note that some people are represented by more fragments of tracks, because we have tracking failures due to long and complete occlusions (person 10 occludes the group). Thus, the lost tracks are reinitialized with a new ID. The associations between the different track fragments are: $(1, 14)$, $(2, 13)$, $(4, 7, 12)$, and $(5, 8)$. Note that the automatic association of IDs is also possible in such scenarios using the person re-identification or re-acquisition methods discussed in the previous chapter. Fig. 5.11 shows that our model is able to detect interactions also when the scene contains multiple groups.

A more sophisticated analysis of accuracy performances of our method is shown in Fig. 5.12 and 5.13. The graphs summarize the group detection accuracy in terms of precision (on the left) and recall (on the right). In the definition of those measurements, we consider as true positive when a group is detected considering

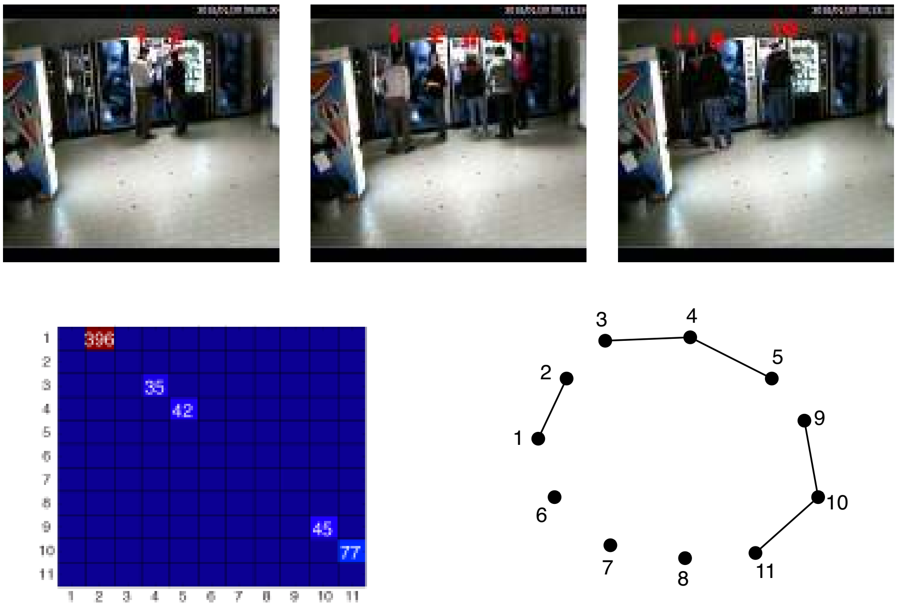


**Fig. 5.11.** Example of cIRPM analysis of sequence $S_{01}$.Three groups (1,2),(3,4,5), and (9,10,11) are detected. One can note that some people are represented by more than one track, since due to severe or complete occlusions the tracks are sometimes lost and need to be reinitialized (e.g. 6,7,8 are reinitialized as 9,10,11, respectively).

all its constitutive members. If a person that belongs to a group is not detected, we have a false negative, and a similar reasoning applies for the false positive.

Fig. 5.12 depicts the statistics as a function of the size of the time interval $T$ frames (x-axis) used to accumulate the IRPM. Each curve corresponds to a value of threshold $Th$ (5, 20, 60 and 100). From this figure, we notice that increasing $T$ gives worse accuracy. Moreover, the peak of each curve depends on both the threshold and the time interval size. We obtain the best performance by setting the $Th$ equal to 20; the peak of this curve corresponds to $T$ equal to 300 frames. Instead, Fig. 5.13 shows the performances increasing the threshold (x-axis) used to detect the groups. Each curve corresponds to a value of $T$ (120, 300, 480, 720, 900, and 1200 frames). The common behavior of all the curves is that increasing and decreasing too much the threshold decreases the accuracy. This analysis confirms that the best performances are given by setting the threshold to 20 and the time interval to 300 frames. When $T$ increases the accuracy drastically decreases and the peak of each curve is shifted, depending by the time interval size.

Intuitively, when the threshold is too low and the time window is too small, our method detects interactions that could contain false positives. Increasing the size of the time window and the threshold permits to average out and cancel out these false positive, because the IRPM becomes more stable. On the other hand, when the threshold is too high, our model is not able to detect interactions, because $cIRPM_t(i, j) > Th$ is zero for each $(i, j)$. To deal with this problem, we could fix the time interval larger. However, in this case, a group interaction interval should be smaller than the time window, and in any case the threshold would result too high to detect groups. For these reasons, precision and recall in Fig. 5.12 and Fig. 5.13, respectively, decrease before and after the optimal setting of the parameters ($Th = 20$ and $T = 300$).

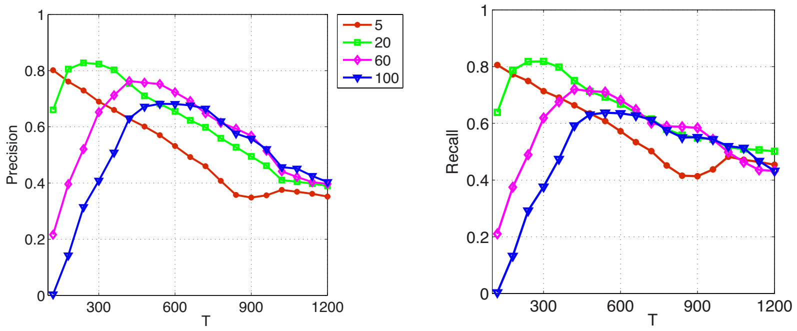


**Fig. 5.12.** Evaluation of precision (left) and recall (right) of the proposed method varying the size of the time interval $T$ (x-axis) used to compute the IRPM. The graph shows one curve for each threshold (5, 20, 60 and 100). The maximum for both the statistics is given by setting $Th = 20$.

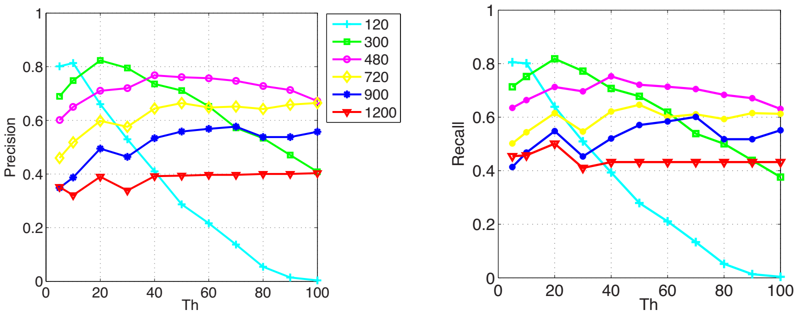


**Fig. 5.13.** Evaluation of precision (left) and recall (right) of the proposed method varying the threshold $Th$ (x-axis) used to detect the groups. The graph shows one curve for each time window (120, 300, 480, 720, 900, and 1200). The maximum for both the statistics is given by setting $T = 300$ and the peak corresponds to $Th = 20$.

### 5.5.2 Interest Map

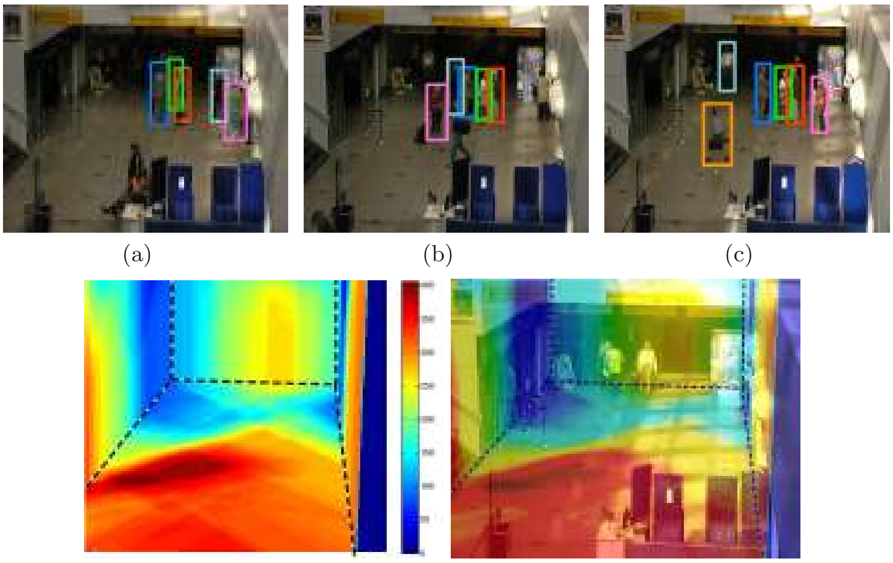
We perform some tests on the publicly available PETS 2007 sequence sets [5], aiming at showing the expressiveness of our framework on widely known and used datasets. Two sequences are considered for the experimental validation, both belong to the S07 dataset depicting an airport area monitoring. The first sequence is captured by Camera 2, the second one is captured by Camera 4.

Figures 5.14(a-c) show the tracking results (bounding-boxes) of three frames of the first considered sequence. Totally, 1 minute of activity has been monitored, tracking continuously 5 people at a time in average. The resulting Interest Map is depicted in Figures 5.14(d,e), superimposed as transparency mask to a frame of the video. The "hottest" area is the one closest to the camera, in the direction of the stairs on the left. Indeed, in the sequence, many people cross that area from right to left. Another interesting area is at the end of the corridor, while the entrance on the left end has never been watched. Finally, the other people detected throughout the sequence are on the right end, going North.

For the second sequence, captured by Camera 4, 1 minute has been monitored, tracking 4 people at a time in average. The SVF analysis produces the results shown in Figure 5.15. In this case, the most seen areas of the parallelepiped (the 3D map) are two (Fig. 5.15(c,d)). The left corner of the parallelepiped is "hot" because most of the people go towards that region of the corridor. The second "hot" area is the area in front of the camera, due to a person loitering there most of the time interval considered. As a comparison we plot together the trajectories of the monitored people (Fig. 5.15(b)). This representation is less meaningful from the point of view of people attention analysis. Our information visualization technique is instead intuitive and it captures in a very simple and richer way where people attention is focused.

## 5.6 Conclusions

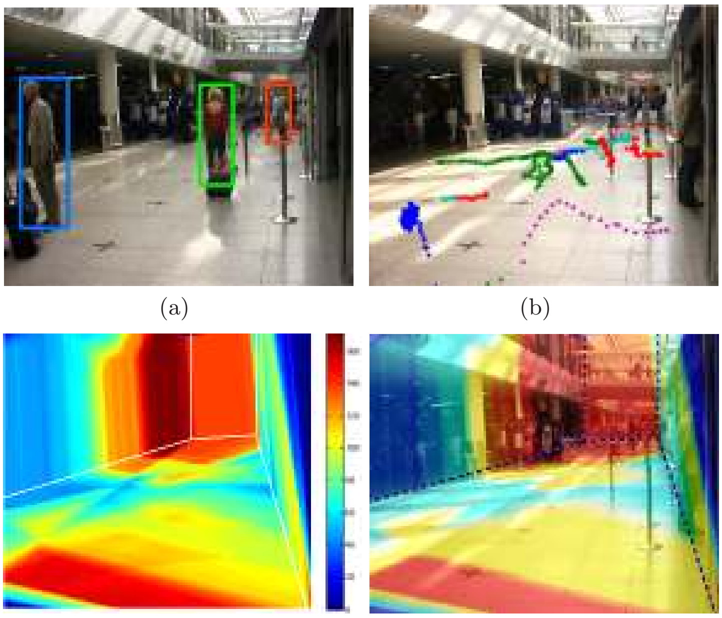
Social signal processing is gaining the researcher interests in the last years because a lot of excellent methods and models have been designed and developed by the machine learning, pattern recognition and computer vision communities. This chapter is a step forward towards the design of social signal process methods and systems for video-surveillance purposes, that are robust and effective. To this purpose, we presented a set of techniques for managing groups and group activities in a principled way, taking into account social psychology aspects that define the human's acting. In this way, we moved from the un-personal objective point of view of the video camera capturing people as they were simple entities, to a new perspective where a subjective viewpoint of the individuals is taken into account. In this scenario, the position of a person is linked with the relative location (and orientation) he/she has with respect to all the other subjects in the scene: actually, what is sensed by the single persons helps more strongly in assessing what he/she is doing with respect to the sterile point of view of a video camera mounted on a wall. This chapter showed how computer vision and social signal processing may collaborate for a new level of the video surveillance research, also depicting the quality of the results such a collaboration can achieve. The next chapter will provide a further analysis in this perspective by investigating how groups evolve over time, introducing the (joint) individual-group tracking.



(a) (b) (c)

(d) (e)

**Fig. 5.14.** (a-c) Tracking results for PETS 2007 S07 camera 2 on three time steps. (d) Resulting Interest Map ("hot" colors represent the areas more frequently observed, and the opposite for the "cold" areas). (e) The same Interest Map superimposed on one frame.



(a)

(b)

(c)

(d)

**Fig. 5.15.** (a) Tracking result for PETS 2007 S07 camera 4 on a single time step. (b) The trajectories of the 4 tracked people estimated throughout the sequence displayed in the same frame. (c) The obtained Interest Map. (d) The same Interest Map superimposed on one frame.

# 6

# Tracking of Social Interactions

Tracking of social interactions is a new emerging research field still in its embryonic state. The goal is to track the detected social interactions over time. In the previous chapter, we presented methods that deal with the detection of social interactions with particular focus on groups. In this chapter, we will face the problem of group analysis over time, that means to perform tracking. A specific case of social interactions tracking is *group tracking* that consists in following tight formations of individuals while they are walking or interacting (Figure 6.1).

One solution for group tracking could consist in extracting some high-level information from the results provided by an individual tracker in a hierarchical way. On the other hand, one can perform both individual and group tracking; the problem takes the name of *individual-group tracking*. One of the major difficulties of the group tracking lies in the high variability of the group entity: splitting, merging, initialization and deletion are frequent events that characterize the life of a group. Think for example when you meet your friends at the restaurant for dinner: some people come by himself/herself, some with a friend but finally all of you will meet up to make up the group of friends that goes to the restaurant together. This process involves the modification of several instances of group to build the final group.

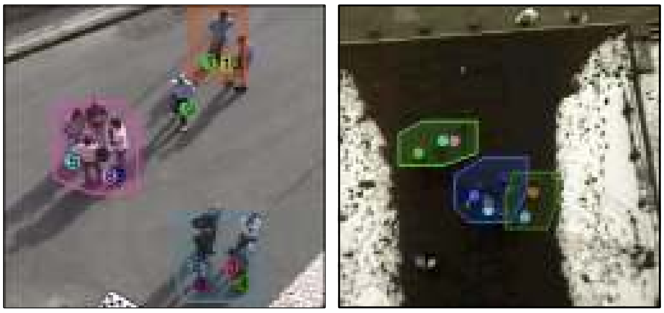


**Fig. 6.1.** Group tracking results (colored convex hulls) of the proposed "Friends meet" dataset (first row) and of BIWI dataset (second row) [191].

## 6.1 Introduction

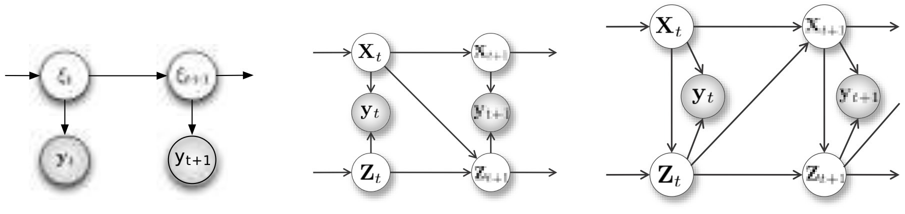
Group tracking is a recent open challenge that can be important in many respects: in computer vision and signal processing, it may help in locating individual targets in the case of missing measurements [186,191]; in surveillance, it may reveal social bonds between people, owing to a high-level scene awareness [55, 62] or increase re-identification rates [275].

In this chapter, two alternative solutions are proposed to cope with the individual-group tracking problem. The first solution relies on the standard *first-detect-then-track* model, where first a detector localizes the groups of interest and then a tracker probabilistically tracks them as an atomic entities. The proposed method is dubbed Collaborative Particle Filtering (Co-PF). Co-PF handles the group and individual tracking separately using particle filtering and then the information is shared in a collaborative way when the individual estimate is available. This helps to keep low the computational burden, but as we will see later Co-PF cannot cope with groups events. For this reason, Co-PF has been deprecated and we prefer to propose an alternative mathematical framework.

The second solution is a *detection-free* model (in terms of groups), that is able to track groups and their change over time in a joint probabilistic framework. The power of this solution is that it is able to deal with group events, such as splitting, merging, initialization and deletion, in a probabilistic way. In contrast, most of the models proposed in literature usually model the group events by heuristic rules, yielding to a scarce generalization. This second method is called DEcentralizEd ParticlE filteR for Joint Individual-Group Tracking (DEEPER-JIGT). This acronym mirrors its potentially deep customizability, that allows to tweak other filtering mechanisms, defining the dynamics of the group given the individual states and vice-versa, how observations are evaluated, etc., as modules of a serial framework.

Let us briefly analyze the proposed probabilistic models for individual-group tracking of Figure 6.2. The first naive solution can be the joint model reported in Figure 6.2(a) ($\xi_t = [\mathbf{X}_t, \mathbf{Z}_t]$ is the joint variable of the individuals state $\mathbf{X}_t$ and the groups state $\mathbf{Z}_t$). Its characteristic relies on the ability of modeling the more general case of joint individual-group tracking. Unfortunately, the model is not suitable for the problem, because inference is intractable and inefficient due to the high-dimensionality of the state space.

The two directions we propose in this chapter go in the direction of making more tractable the problem. In particular, Co-PF (the model in Figure 6.2(b)) approximates the individual-group tracking problem into two *independent* problems: individual tracking and group tracking. The state estimate is performed independently and afterward the information flows down through the probabilistic link between $\mathbf{X}_t$ and $\mathbf{Z}_{t+1}$ also called collaborative link. However, the hierarchical nature of the individual-group tracking problem makes this solution not able to deal with group events. On the other hand, DEEPER-JIGT (the model in Figure 6.2(c)) exploits the joint nature of the problem in order to split it into two *nested* subproblems. This is the basic idea of the Decentralized Particle Filter (DPF) [52] that we employ in order to perform inference in the proposed model. The DPF factorizes the joint individual-group state space in two dependent subspaces, so that it is



(a) Joint state space model (b) Co-PF model (c) DEEPER-JIGT model

**Fig. 6.2.** Different models for filtering. (a) Classical particle filtering in the joint state space $\xi_t = [\mathbf{X}_t, \mathbf{Z}_t]$. (b) Collaborative particle filter in the independent state spaces (one for individual tracking and one for group tracking). (b) DEEPER-JIGT, *i.e.*, state decomposition of the joint state space with DPF.

possible to model the single individuals, and the groups given the knowledge of the individuals. It is worth noting that inference is hard and not as efficient as the disjoint model even using DPF, but the introduction of two approximations through importance sampling will make it more computationally feasible than the complete joint model.

Both Co-PF and DEEPER-JIGT have their strengths and drawbacks. The main advantages of the Co-PF are that: 1) it is able to cope with partial occlusions between individuals and between groups, because it relies on the HJS filter, 2) it is computational efficient, because we deal with two joint state spaces (individuals and groups) that can be further divided in independent sub-spaces through the HJS filter, 3) the probabilistic link between the state spaces is estimated a-posteriori, when $\mathbf{X}_t$ and $\mathbf{Z}_{t+1}$ have been estimated. However, Co-PF is not able to deal properly with groups events. It relies on the assumption that once a group is initialized, the tracker estimates its trajectory as an atomic entity. This drawback ponderously penalizes the Co-PF, because generally groups have variable dynamics.

Many interesting qualities can be ascribed to DEEPER-JIGT: 1) the absence of heuristics to handle group events: they are all governed by probability distributions whose parameters can be learned from training data. 2) DEEPER-JIGT updates the group information in an online fashion, where all the tracking history of the individuals is intrinsically exploited by its composite filtering mechanism. This is in contrast with the widely adopted individual-based analysis methods, where groups are estimated by grouping together short individual trajectories (tracklets) collected beforehand, whose length is typically a critical parameter to be tuned [51, 64, 93, 163, 169, 191, 270]. 3) Finally and more importantly, DEEPER-JIGT allows to understand in a quantitative way how much the modeling of the single targets helps the group tracking and *vice-versa*, suggesting that a joint treatment is beneficial for both worlds.

Our proposals has been evaluated on both simulated and different real scenarios. Going from the Co-PF to the DEEPER-JIGT, we noticed that no dataset with groups events was available in literature, even if it is common to see these events in a camera-monitored scenario. In fact, most of the public datasets have been built for the person detection and tracking tasks, this means that usually

| Group-based class | Individual-based class | Individual-group class |
|---|---|---|
| [79, 94, 261] | [64, 163, 169, 191] | **Co-PF**, **DEEPER-JIGT** |
| [144, 155] | [51, 93, 270] | [98, 168, 186] |

**Table 6.1.** Taxonomy of the existing group tracking methods.

the monitored areas are transition zones. Instead, groups events occur in area where people meet and stay for short or long periods. For this reason we provide a novel benchmark dataset, named *Friends Meet*. The dataset has beed recorder in a zone where people meet and share their break time. Thus, it usually happens that groups pops out, break out, enter and exit from the scene. Moreover, since group tracking evaluation measures do not exist, the existing individual tracking measures [38, 228] have been adapted here to handle groups.

The rest of the chapter is organized as follows. In Section 6.2, a novel taxonomy illustrates the literature on group tracking. In Section 6.3 and 6.4, we present the proposed models to deal with individual-group tracking: the Co-PF and the DEEPER-JIGT, respectively. A thorough experimental section is reported in Section 6.5, and, finally, Section 6.6 concludes the chapter and envisages the future work.

## 6.2 Related Work

The recent literature on group analysis can be partitioned in three categories as shown in Table 6.1: 1) the *group-based* class of techniques where groups are treated as genuine atomic entities without the support of individual tracks statistics; 2) the *individual-based* class, where group descriptions are built by associating individuals tracklets that have been calculated beforehand (typically, with a time lag of few seconds); 3) the *individual-group* class, where group tracking and individual tracking are performed simultaneously. The proposed methods are in this last, more general category.

The group-based approaches are proposed especially when the scene is highly cluttered so that individual tracking cannot be performed, and the detection of the single targets is unreliable. They assume the groups as nonparametric regions [261], Gaussian-shaped distributions [79, 94], clusters over graphs structures [144], textures [155]. As tracking engines, they employ standard approaches, such as Kalman filtering [79, 94], probability hypothesis density filter [261], multi-hypothesis filtering [144], or particle filtering [155].

In the individual-based category, compact regions are classified as different entities, including groups or persons, exploiting a set of heuristics [64,169]. In [64], people that stand close for a while are joined into groups through a connection graph built exploiting heuristics on the moving regions. More principled approaches employ generative modeling [191], discriminative reasoning [270], weighted connection graphs [51] and bottom-up hierarchical clustering [93]. An interesting by-product is presented in [163] where group tracking is employed for facing individual occlusions.

Both the group-based and individual-based classes have drawbacks. Group-based techniques are limited by the fact that individual trajectories are not analyzed, reducing in simplistic models. In the individual-based approaches, the performance is very dependent on the quality of the individual tracklets; more important, groups are seen as mere consequential events of the behavior of the single targets, whereas it is widely known in sociology that groups exert important influence on the acting of the singles.

Individual-group techniques deal with individuals and groups simultaneously. Many of them keep the structure of a graph in which connected components correspond to groups of individuals: in [186], stochastic differential equations are embedded in a Markov-Chain Monte Carlo (MCMC) framework, implementing a probabilistic transition model for the group dynamics. The problem of MCMC is that, in its basic form, does not scale efficiently in high-dimensional state spaces. Lately, in [98], a similar framework has been augmented by considering inter-group closeness and intra-group cohesion. In both cases, experiments with few targets (up to 4) have been presented. A hierarchical structure (two levels) for tracking that uses a physically-based mass-spring model is proposed in [168]: the first level deals with individual tracking, and the second level tracks individuals that are spatially coherent. Similarly in principle, in Co-PF, two processes are involved: the group process considers groups as atomic entities. The individual process captures how individuals move, and revises the group posterior distribution. However, both of them do not considers split and merge events.

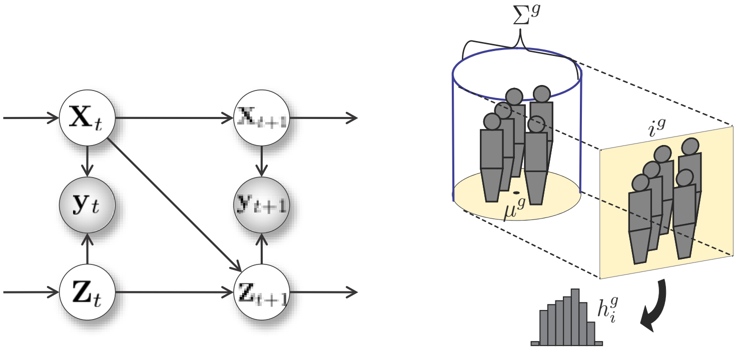
Also DEEPER-JIGT lies in this last category, differing from the state of the art in many aspects, primarily in the filtering mechanism which was inspired by the decentralized particle filter [52]. Moreover, as we will show in the following sections, DEEPER-JIGT allows to simultaneously deal with merge and split events and with a varying number of individuals and groups. Most important, it allows to understand through quantitative measures the effectiveness of the collaboration of individual and group processes for tracking, promoting the latter category of tracking approaches as the most promising one.

## 6.3 Collaborative Particle Filtering

Collaborative particle filtering is a technique to probabilistically link two state spaces. The model depicted in Figure 6.3(a) is made by two processes, in the specific case, two HMMs $\mathbf{X}_t$ and $\mathbf{Z}_t$ that share the same observations, $\{\mathbf{y}_t\}$. This means that the two processes evaluate the scene under two different resolutions. When observing the same thing at two different resolutions, there exists a dependence on what they observe and what they estimate. In other words, the estimates of $\mathbf{X}_t$ and $\mathbf{Z}_t$ have to be correlated in some way.

In Co-PF, we explicit the dependence between the states of the system by adding the probabilistic link from $\mathbf{X}_t$ to $\mathbf{Z}_{t+1}$. Notice that we avoid to add the link between $\mathbf{X}_t$ to $\mathbf{Z}_t$ because inference becomes harder. In our choice, first we can estimate $\mathbf{X}_t$ to $\mathbf{Z}_t$ and then we use the information of $\mathbf{X}_t$ for the next time step. In this way, the estimation method remains online and simple.

In individual-group tracking, we split the problem into individual tracking $\mathbf{X}_t$ (lower resolution) and group tracking $\mathbf{Z}_t$ (higher resolution). In the following, we will see how the proposed model is defined for our problem and which approximations we need, to perform inference on it.



(a) Co-PF (b) Rendering function

**Fig. 6.3.** Collaborative PF idea and group rendering.

Both individual tracking and group tracking are modeled by an HJS filter [142] (details in Section 2.2.1), in order to enable the model to deal with occlusions between people and between groups. It is worth noting that the model is not constrained by the chosen filter, but the strategy can be applied also to other particle filters defined in the joint state space.

Each individual state is modeled as an elliptical shape on the ground plane, *i.e.*, $\mathbf{x}_t^k = \langle \mu^k, \Sigma^k \rangle$, where $\mu^k$ is the position of the individual on the ground plane, $\Sigma^k$ is a covariance that measures the occupancy of the body projected on the ground plane. The covariance matrix is set to be constant for each individual as approximation of the human silhouette, *i.e.*, $\forall k = \{1, 2, \ldots, K\}, \Sigma^k = \Sigma$. Note that, we are going to need the calibration parameters to use this model of group.

As for group tracking, we denote the $g$th group as $\mathbf{z}^g = \langle \mu^g, \Sigma^g \rangle$, where $\mu^g$ is the 2D position on the ground floor of the $g$th group and $\Sigma^g$ is the covariance matrix that approximates the projection of its shape on the floor (Fig. 6.3(b)). This time we cannot fix $\Sigma^g$ for each group because each group has its own shape. The choice of an ellipse for modeling the floor projection of a group is motivated from a sociological point of view, exploiting proxemics notions that describe a group as a compact closed entity [103]. However, we will see in the experiments that this is not the optimal choice of the state space definition. For this reason, we propose a more compact definition in the next section: we will use labels and dynamics over labels for DEEPER-JIGT that also enables us to deal with group events.

The posterior distribution of the $g$th group in group tracking follows the standard Bayesian recipe already seen in Chapter 2 (Eq. 2.1):

$$p(\mathbf{z}_{t+1}^g | \mathbf{y}_{1:t+1}) \propto p(\mathbf{y}_{t+1} | \mathbf{z}_{t+1}^g) \int p(\mathbf{z}_{t+1}^g | \mathbf{z}_t^g)\, p(\mathbf{z}_t^g | \mathbf{y}_{1:t})\, d\mathbf{z}_t^g . \tag{6.1}$$

Also in this case, the filtering mechanism is fully specified by an initial distribution $p(\mathbf{z}_0^g)$, a dynamical model $p(\mathbf{z}_{t+1}^g|\mathbf{z}_t^g)$, and an observation model $p(\mathbf{y}_{t+1}|\mathbf{z}_{t+1}^g)$, that are opportunely sampled, generating a set of *group* particles. The same Bayesian formulation is also run in parallel for the individual tracker, just replace $\mathbf{z}^g$ with $\mathbf{x}^k$ in Eq. 6.1.

In the next two sections, we will first define the chosen group dynamical and observation models and then we will derive the mathematical formulation of the collaborative link.

### 6.3.1 Group Distributions

The dynamical model $p(\mathbf{z}_{t+1}^g|\mathbf{z}_t^g)$ is define in the same way of Eq. 2.5 using the HJS filter. Therefore, if we define the joint group variable as $\mathbf{Z}_t = \{\mathbf{z}_t^1, \mathbf{z}_t^2, \ldots, \mathbf{z}_t^G\}$, with $G$ the number of groups in the scene, then the joint dynamical model can be approximated as $p(\mathbf{Z}_{t+1}|\mathbf{Z}_t) \approx p(\mathbf{Z}_{t+1}) \prod_g q(\mathbf{z}_{t+1}^g|\mathbf{z}_t^g)$. The function $q(\mathbf{z}_{t+1}^g|\mathbf{z}_t^g)$ is modeled by considering the nature of $\mathbf{z}_{t+1}^g = \langle \mu^g, \Sigma^g \rangle$, that is a Gaussian distribution. Since $\mu^g$ and $\Sigma^g$ are completely independent variables, we can define the dynamics disjointly. For the mean of the Gaussian $\mu^g$, we assume a linear motion, perturbed by white noise with parameter $\sigma_\mu$ like in the standard models we have discussed about in Chapter 2.

The dynamics of the covariance matrix $\Sigma^g$ is defined by the perturbation of the principal axes of the Gaussian. Since the matrix $\Sigma^g$ is always a real nonsingular symmetric matrix with orthonormal eigenvectors, it can be decomposed as $\Sigma_g = V \Lambda V^T$, where $\Lambda = \mathrm{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_E)$ contains the eigenvalues $\{\lambda_i\}_{i=1}^E$ and $V = [\mathbf{v}_1 \mathbf{v}_2 \ldots \mathbf{v}_E]$ contains the eigenvectors $\{\mathbf{v}_i\}_i^E$ (in our case $E = 2$, *i.e.*, the ground floor position). This formulation makes easier the perturbation of the principal axes, by simply perturbing the direction of the eigenvectors and the module of the eigenvalues preserving the constrains of the eigenvectors. In fact, we rotate the principal axes with respect to the $z$ axis by an angle $\theta$, by applying the rotation matrix $R(\cdot)$ to the eigenvectors:

$$V' = [R(\mathcal{N}(\theta, \sigma_\theta))\,\mathbf{v}_1,\; R(\mathcal{N}(\theta, \sigma_\theta))\,\mathbf{v}_2] \tag{6.2}$$

in this way they are still orthonormal. Then, we modify the eigenvalues by varying the amplitude of the principal axes:

$$\Lambda' = \begin{bmatrix} \mathcal{N}(\lambda_1, \sigma_\lambda) & 0 \\ 0 & \mathcal{N}(\lambda_2, \sigma_\lambda) \end{bmatrix} \tag{6.3}$$

Note that instead of using the exact values of $\theta$ and $\lambda$, we use a version with Gaussian noise with $\sigma_\theta$ and $\sigma_\lambda$ as variances, assuming a linear dynamics with Gaussian noise in the spectral space. The matrixes $V'$ and $\Lambda'$ are then used to recompose the new hypothesis $\Sigma_g' = V' \Lambda' V'^T$, that will represent a new perturbed elliptical shape. The advantage of this representation is that we can actually model the 3D spatial displacement of the group. However, the joint state space $\mathbf{Z}$ will become huge. The HJS filter helps us to perform inference in this huge state space using approximations. This representation is not the best if the problem is strictly joint. A better representation is the one based on labels that is proposed in the next section.

The dynamics prior $p(\mathbf{Z}_{t+1})$ implements an exclusion principle that cancels out inconsistent hypotheses. In other words, two groups' hypotheses, that are close and partially overlapped, will be rejected. We employ a Markov Random Field learned via Belief Propagation as in Section 2.2.1, where each node represents a group hypothesis and the weight of each link is computed considering the overlapping area of the two hypotheses. In practice, the weight is defined as

$$\frac{\mathrm{Over}(\mathbf{z}^g, \mathbf{z}^h)}{\mathrm{Area}(\mathbf{z}^g) + \mathrm{Area}(\mathbf{z}^h)},$$

where $\mathrm{Area}(\mathbf{z}^g)$ is the area of the ellipse projected in the ground plane (see Figure 6.3(b)) and $\mathrm{Over}(\mathbf{z}^g, \mathbf{z}^h)$ is the overlapping area of the projected ellipses of the $g$-th and $h$-th group.

The observation model $p(\mathbf{y}_{t+1}|\mathbf{z}^g_{t+1})$ is derived as Eq. 2.6. In order to easily evaluate an observation $\mathbf{y}_{t+1}$, we employ a *rendering function* $f(\cdot)$ that maps a state in a convenient feature space. The idea is depicted in Figure 6.3(b): The rendering function projects the elliptical volume in the image using the calibration parameters. The projection area will be a rectangle from which is possible to extract features (we use color histograms). The observation model is defined in Gibbs form as we have already seen in the previous chapters:

$$p(\mathbf{y}_{t+1}|\mathbf{z}^g_{t+1}) = e^{(-\lambda_z d(f(\mathbf{z}^g_{t+1})), f(\tau^g_{t+1}))},$$

where $d$ is a distance between color histograms (Bhattacharyya distance in our experiments), $\mathbf{z}^g_{t+1}$ is a group hypothesis and $\tau^g_{t+1}$ is the template of the group. The template is built when the new group is detected in the scene, its centroid $\mu^g$ and occupancy area $\Sigma^g$ are estimated, forming the initial state $\mathbf{z}^g_{t+1}$. We build a volume of height 1.80m upon the area $\Sigma^g$, in order to surround the people of the group. The template has also to be modified in order to deal with deformations. Thus, the rendering function takes the template and deforms it opportunely by a re-scaling, considering the $\mu'^g$, and by a shearing, taking into account the deformation resulted by the perturbation of the covariance matrix $\Sigma'^g$.

The joint observation model $p(\mathbf{y}_{t+1}|\mathbf{Z}_{t+1})$ considers all the groups present in a scene, taking into account what part of the group $\mathbf{z}^g_{t+1}$ is seen with respect to the remaining groups $\mathbf{Z}^{\neg g}_{t+1}$. In other words, an occlusion map is built for each group. A problem in our definition is that we assume a group as a rigid solid shape and this permits to model inter-group occlusions, but it does not model intra-group occlusions (i.e., persons of a group that mutually occlude each other). This leads to tracking applications where a strong intra-group occlusion causes the loss of that group.

Choosing this group observation model is actually not the best way to model the group appearance. The appearance template of a group is not well defined in general, because many components of a group are usually occluded, also when the group is initialized and the method acquire the first template. Moreover, since people in a group continues to change their relative position, the appearance dynamic vary in a highly non-linear way. Combined to the fact that learning the template is usually hard, this makes the model not so robust. This observation motivates the choice of labels as group description in DEEPER-JIGT. When using labels, we are not going to need any template and hence any template update algorithm for groups.

### 6.3.2 The Collaborative Probabilistic Link

Co-PF solves the intra-group occlusions problem, and permits a fine estimate of the whereabouts of a scene. It basically injects the information collected by the individual tracker into the group tracker. Considering the filtering expression in Eq. 6.1, the fusion occurs on the posterior at time $t$:

$$p(\mathbf{z}_t^g|\mathbf{y}_{1:t}) = \int p(\mathbf{z}_t^g, \mathbf{X}_t|\mathbf{y}_{1:t})\, d\mathbf{X}_t = \int p(\mathbf{z}_t^g|\mathbf{X}_t, \mathbf{y}_{1:t})\, p(\mathbf{X}_t|\mathbf{y}_{1:t})\, d\mathbf{X}_t. \tag{6.4}$$

The first term of Eq. 6.4 is the core of our approach as it revises the group posterior distribution at time $t$, considering the joint state of the individuals by simply marginalizing over $\mathbf{X}_t$. In this way, the second term (the posterior at time $t$ of the individual tracker) may be considered as a weight that mirrors the reliability of the individual states.

A convenient way to model distributions conditioned on multiple events is that of the Mixed-memory Markov Process (MMP) [218], that decomposes a structured conditioned distribution as a convex combination of pairwise conditioned distributions. This leads to the following approximation:

$$p(\mathbf{z}_t^g|\mathbf{X}_t, \mathbf{y}_{1:t}) \approx \alpha\, p(\mathbf{z}_t^g|\mathbf{X}_t) + (1-\alpha)\, \widetilde{p}(\mathbf{z}_t^g|\mathbf{y}_{1:t}), \tag{6.5}$$

where $\alpha \in [0, 1]$. Considering Eq. 6.5, we can rewrite Eq. 6.4 as:

$$p(\mathbf{z}_t^g|\mathbf{y}_{1:t}) \approx \alpha \int p(\mathbf{X}_t|\mathbf{y}_{1:t})\, p(\mathbf{z}_t^g|\mathbf{X}_t)\, d\mathbf{X}_t + \tag{6.6}$$

$$(1-\alpha)\; \widetilde{p}(\mathbf{z}_t^g|\mathbf{y}_{1:t}) \underbrace{\int p(\mathbf{X}_t|\mathbf{y}_{1:t})\, d\mathbf{X}_t}_{=1}. \tag{6.7}$$

At this point, it is easy to realize that $p(\mathbf{z}_t^g|\mathbf{y}_{1:t})$ becomes a combination of the natural group posterior and a marginalization of the *linking* probability $p(\mathbf{z}_t^g|\mathbf{X}_t)$, that relates the $g$th group to the individuals, weighted by the individual tracking posterior $p(\mathbf{X}_t|\mathbf{y}_{1:t})$. In other words, the group posterior is revisited by injecting in a principled way the information on the individuals, conveyed selectively by $p(\mathbf{z}_t^g|\mathbf{X}_t)$.

In the proposed framework, we can exploit the advantages of the formulation of the HJS filter in order to compute $p(\mathbf{X}_t|\mathbf{y}_{1:t})$. Alternatively, one can define it directly using its joint formulation. Therefore, Eq. 2.4 helps us to decompose the posterior of all the individual states 6.6 as follows:

$$p(\mathbf{X}_t|\mathbf{y}_{1:t}) \approx \prod_{k=1}^{K} p(\mathbf{x}_t^k|\mathbf{y}_{1:t}). \tag{6.8}$$

To approximate Eq. 6.6, we can use the particle set $\{(\mathbf{x}_t^{(n)}, w_t^{(n)})\}_{n=1}^N$ provided by the individual tracker. The idea is to sample a new particle set from each $k$-th distribution $p(\mathbf{x}_t^k|\mathbf{y}_{1:t})$ and then we compute $p(\mathbf{z}_t^g|\mathbf{X}_t)$ for each particle. Then, a new set of weighted particles that approximates Eq. 6.6 is built. In this way, the computational cost of our method is kept polynomial in the number of objects and particles.

The linking probability $p(\mathbf{z}_t^g|\mathbf{X}_t)$ is factorized as an MMP as follows:

$$p(\mathbf{z}_t^g|\mathbf{X}_t) \approx \sum_{k=1}^{K} \beta^{k,g}\, p(\mathbf{z}_t^g|\mathbf{x}_t^k) \tag{6.9}$$

$$\propto \sum_{k=1}^{K} \beta^{k,g}\, p(\mathbf{x}_t^k|\mathbf{z}_t^g)\, p(\mathbf{z}_t^g) \tag{6.10}$$

where $\beta^{k,g} > 0\ \forall k, g$ and $\sum_k \beta^{k,g} = 1$. Each term of the sum in Eq. 6.9 represents the probability that the $g$-th group $\mathbf{z}_t^g$ contains the $k$-th target $\mathbf{x}_t^k$. In Eq. 6.10, we simplifies the formulation by employing the Bayes rule, where $p(\mathbf{x}_t^k|\mathbf{z}_t^g)$ defines the *linking* likelihood that each individual state $\mathbf{x}_t^k$ is a subpart of $\mathbf{z}_t^g$.

Finally, the entire model ends up in the definition of three components: the linking likelihood $p(\mathbf{x}_t^k|\mathbf{z}_t^g)$, the group prior $p(\mathbf{z}_t^g)$ and the weights $\beta^{k,g}$. The linking likelihood is defined here in terms of three components: 1) the appearance similarity, 2) the dynamics consistency, and 3) the group membership. The appearance similarity is encoded by the Bhattacharyya distance between the HSV histograms of the two renderized entities: $d_{\mathrm{HSV}}(f(\mathbf{z}_t^g), f(\mathbf{x}_t^k))$. The dynamics consistency rewards the person state whose motion component is similar to that of the group. In practice, we check the 2D displacement on the floor by calculating $d_{\mathrm{dir}}(\mathbf{z}_t^g, \mathbf{x}_t^k) = |1 - |\mathrm{dir}(\mathbf{z}_t^g) - \mathrm{dir}(\mathbf{x}_t^k)|/\pi|$, where $\mathrm{dir}(\cdot)$ gives the direction (an angle) of the person or group. Finally, the group membership evaluates the spatial proximity of the person state and of the group state:

$$d_{\mathrm{mbr}}(\mathbf{z}_t^g, \mathbf{x}_t^k) = \begin{cases} 1 \text{ if } & \mathbf{x}_t^k \in \mathbf{z}_t^g \\ 0 \text{ otherwise} \end{cases} \tag{6.11}$$

where the membership operator $\in$ controls if the $k$-th person position is inside the $g$-th group ellipse. Therefore, $p(\mathbf{x}_t^k|\mathbf{z}_t^g) \propto d_{\mathrm{HSV}}(\mathbf{z}_t^g, \mathbf{x}_t^k) \cdot d_{\mathrm{dir}}(\mathbf{z}_t^g, \mathbf{x}_t^k) \cdot d_{\mathrm{mbr}}(\mathbf{z}_t^g, \mathbf{x}_t^k)$, because the distances are normalized between 0 and 1. The coefficients $\beta^{k,g}$ express a linking preference that an object belongs to a group, and are left here as uniform, *i.e.*, $\beta^{k,g} = 1/G$. In order to compute the integral in Eq. 6.6, the evaluation of $p(\mathbf{x}_t^k|\mathbf{z}_t^g)$ is done for each particle of the group $g$ and for each ones of the person $k$. Then, the values are summed over the person particles, performing the Monte Carlo approximation. The same procedure is employed for computing $\beta$s' value, however we also sum over the group particles in order to obtain a single $\beta$ for each $k$ and $g$.

Finally, the prior $p(\mathbf{z}_t^g)$ discards the biggest and the smallest group hypotheses, rejecting the particles in which the size of the group is below a threshold $\tau_b$ or above a threshold $\tau_a$.

An example that explains the power of our formulation can be represented by an intra-group occlusion in the $g$-th group at time $t$, which is very common due to the dynamic nature of a group of moving people. Let $\mathbf{x}_t^k$ a target of the group $\mathbf{z}_t^g$ that vanishes as occluded by the remaining individuals of that group. The group posterior $p(\mathbf{z}_t^g|\mathbf{y}_{1:t})$ will not be very high, for the limits of the visual, rigid, group representation. However, the individual tracker will "understand" that $\mathbf{x}_t^k$ is occluded, producing a high value for $p(\mathbf{x}_t^k|\mathbf{y}_{1:t})$. This probability value will flow through $p(\mathbf{z}_t^g|\mathbf{X}_t)$, which is high because, even if occluded, the position and the velocity of $\mathbf{x}_t^k$ are correctly estimated by the individual tracker, and will give a high linking likelihood. This will reinforce the final estimation of the hybrid posterior for $\mathbf{z}_t^g$, thus enabling to estimate the subsequent group sample set in a more robust way.

## 6.4 Decentralized Particle Filter for Joint Individual-Group Tracking

In the previous section, we have described a model that factorizes the joint state space in two independent state spaces with a probabilistic link between the posterior distribution over the two spaces. This has introduced additional approximations to deal with inference in the model, as we have seen in Section 6.3.2. In this section, we propose a new model that deals with the joint formulation of the state space.

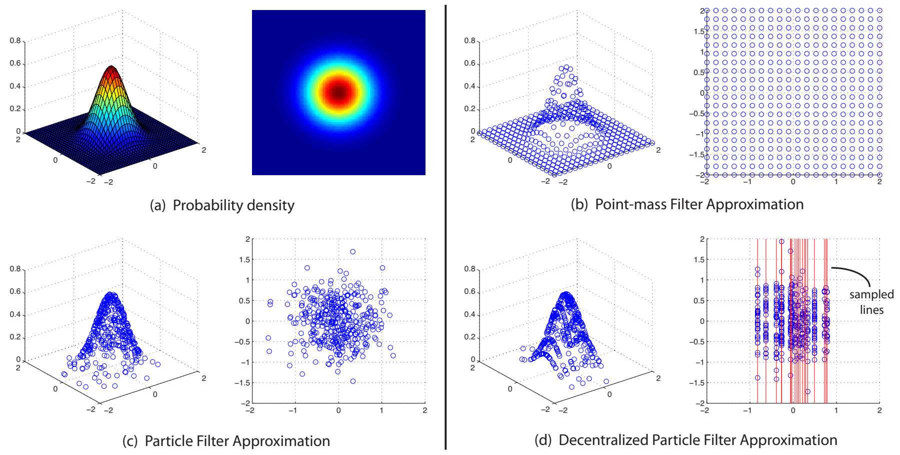
Figure 6.5 shows the proposed factorized model, that we have called DEcentralizEd ParticlE filteR for Joint Individual-Group Tracking (DEEPER-JIGT). The main idea is that the joint model can be factorized in a way that inference is split in two nested subproblems that correspond to the state spaces for individuals and for groups. The two subproblems are solved one after the other using sequential Monte Carlo approximation methods, such as importance sampling. Once estimated the posterior distribution in one part of the state space, we use that information to estimate the posterior distribution over the other part of the state space. The combination of the two distributions enables us to perform approximate inference in the joint state space. A key role is played by the Decentralized Particle Filter (DPF) proposed in [52].

Section 6.4.1 and Section 6.4.2 are devoted to give an intuitive preview of the DPF and a review the DPF with particular focus on how it is used in this work, respectively. Then, Section 6.4.3 describes how we instantiate the model for JIGT, that is, every probability distribution is defined.

### 6.4.1 Intuitive Preview of The Decentralized Particle Filter

The DPF has been recently presented and described in [52]. We follow the example presented by the authors of [52], that gives an intuition on how the DPF works in practice, because there is the risk that the reader will be distracted by the high load of formulae presented in the next sections.

Let us assume for the sake of visualization to have a two-dimensional state space with components $x$ and $z$. The probability distribution over $x$ and $z$ will be a surface over the $(x, z)$ plane; a simple example is depicted in Figure 6.4(a), that is, a Gaussian distribution. The goal of filtering is to estimate the surface. One naive solution is given by the point-mass filter: it evaluates the values on a fixed grid (Figure 6.4(b)). A more interesting solution is provided by the classical



(a) Probability density

(b) Point-mass Filter Approximation

(c) Particle Filter Approximation

(d) Decentralized Particle Filter Approximation

**Fig. 6.4.** Example of bivariate probability distribution. (a) The standard surface of a Gaussian distribution over the plane (3D visualization and from the top of the plane). (b) The point-mass filter approximation on a fixed grid. (c) The classical particle filter approximation with random particles. (d) The decentralized particle filter approximation where first the red lines are sampled and then particles are generated on the lines accordingly some sampling distribution.

particle filter: it throws points following a certain proposal distribution and let them move to important regions (Figure 6.4(c)). The third solution is the DPF way: first choose a set of parallel lines in the $x$ axis and then throw points at random with the only restriction that they have to stay along the chosen parallel lines (Figure 6.4(d)).

### 6.4.2 The Decentralized Particle Filter

The DPF [52] addresses the classical non-linear discrete-time system of Figure 6.2(a)

$$\xi_{t+1} = f_t(\xi_t, \eta_t), \quad \mathbf{y}_t = h_t(\xi_t, \eta_t^y) \tag{6.12}$$

where $\xi_t$ is the state of the system at time $t$, $\mathbf{y}_t$ is the observation or measurement, $\eta_t$ and $\eta_t^y$ are independent non-Gaussian noises, and $f_t$ and $h_t$ are nonlinear functions . The goal of DPF is that of recursively estimating the posterior distribution $p(\xi_t|\mathbf{y}_{0:t})$ through a *decomposition* of $\xi_t$ in two (or more) subspaces, *i.e.*, $\xi_t = [\mathbf{X}_t, \mathbf{Z}_t]^T$. Therefore, Eq. 6.12 can be reformulated as:

$$\begin{aligned}
\mathbf{X}_{t+1} &= f_t^x(\mathbf{X}_t, \mathbf{Z}_t, \eta_t^x), \\
\mathbf{Z}_{t+1} &= f_t^z(\mathbf{X}_t, \mathbf{Z}_t, \eta_t^z), \\
\mathbf{y}_t &= h_t(\mathbf{X}_t, \mathbf{Z}_t, \eta_t^y).
\end{aligned}$$

The main factorization that enables us to split the original problem in two nested subproblems is:

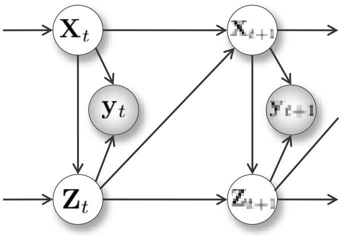


**Fig. 6.5.** State decomposition of the joint state $\xi_t = [\mathbf{X}_t, \mathbf{Z}_t]$ with DPF.

$$p(\mathbf{Z}_t, \mathbf{X}_{0:t}|\mathbf{y}_{0:t}) = p(\mathbf{Z}_t|\mathbf{X}_{0:t}, \mathbf{y}_{0:t})\, p(\mathbf{X}_{0:t}|\mathbf{y}_{0:t}). \tag{6.13}$$

Note that this formulation is not an approximation as in Co-PF. The factorization enables DPF to circumvent both the inefficiency and ineffectiveness of the classical particle filtering when dealing with large sized $\xi_t$'s [52]. Clearly, the two subproblems become the estimate of the two nested distributions: 1) $p(\mathbf{X}_{0:t}|\mathbf{y}_{0:t})$ and 2) $p(\mathbf{Z}_t|\mathbf{y}_{0:t}, \mathbf{X}_{0:t})$. Such distributions are analyzed in a serial way, detailed in Algorithm 4. The underlying idea is that $p(\mathbf{X}_{0:t}|\mathbf{y}_{0:t})$ explains a subspace of the joint space (related to $\mathbf{X}$), and that knowledge is injected into the estimation of $\mathbf{Z}$ in $p(\mathbf{Z}_t|\mathbf{y}_{0:t}, \mathbf{X}_{0:t})$ through the conditional chain rule. More in details, DPF performs two numerical approximations by importance sampling, explaining both terms of Eq. 6.13 at time $t$ (steps 1-3), then moving to step $t+1$ (steps 4-7). The distributions highlighted in gray will be explained in the next section. Distributions with subscripts (*e.g.*, $p_{N_z}$) are approximated by samples, and are not described in parametric form.

In Step 1, the standard importance sampling formulation (*Observation* $\cdot$ *Dynamics*)/(*Proposal distribution*) (see Equation 2.2) is applied for approximating $p(\mathbf{X}_{0:t}|\mathbf{y}_t)$. The difference with the standard framework lies in the term $\mathbf{y}_{0:t}$, whose formal presence is motivated by a mathematical derivation discussed in [52]. Intuitively, the conditioning of $\mathbf{y}_{0:t}$ injects the knowledge acquired by explaining $\mathbf{y}$ in the $\mathbf{Z}$ subspace at time $t$. This highlights the bidirectional relationship of the processes that analyze $\mathbf{X}$ and $\mathbf{Z}$ because, during the same time step, operating on $\mathbf{X}$ helps in better defining $\mathbf{Z}$, and across subsequent time steps, operating on $\mathbf{Z}$ helps $\mathbf{X}$. Step 2 is a classical re-sampling, that regularizes the distributions of the samples (their variance being diminished). Step 3 approximates $p(\mathbf{Z}_t|\mathbf{X}_{0:t}, \mathbf{y}_{0:t})$ by importance sampling, assuming the dynamics equal to the proposal (so dividing by one). After that, predictions for time $t+1$ are made. As for the previous time step, the $\mathbf{X}$ subspace is first analyzed, sampling particles according to a given dynamics $\pi(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}^{(i)}, \mathbf{y}_{0:t})$. The information encoded in that sample set is plugged into the importance sampling approximation of the posterior $p(\mathbf{Z}_t|\mathbf{X}_{0:t+1}, \mathbf{y}_{0:t})$ (Step 5), yielding to a second resampling step (Step 6) and to the final sampling of $\mathbf{Z}$ at time $t+1$ (Step 7).

In the original paper [52], the approach was tested with simulations on 2D (4D) points, where $\mathbf{X}$ and $\mathbf{Z}$ lies in two $\mathbb{R}^2$ ($\mathbb{R}^4$) subspaces. In our case, we are dealing with a much more intriguing and complex problem, where the subspaces have completely different meaning, other than being high-dimensional. In particular, $\mathbf{X}$

**Algorithm 4**: The DPF algorithm [52]. INPUT: samples $\{\mathbf{X}_{0:t}^{(i)}\}_{i=1,\ldots,N_x}$, samples $\{\mathbf{Z}_{0:t}^{(i,j)}\}_{i=1,\ldots,N_x,j=1,\ldots,N_z}$. The superscripts $(i,j)$ mean that for each $i$ particle generated for describing $\mathbf{X}$ we have $N_z$ particles for describing $\mathbf{Z}$. OUTPUT: importance sampling approximations of $\mathbf{X}_{t+1}$, $\mathbf{Z}_{t+1}$.

```
1. Approximation of $p(\mathbf{X}_{0:t}|\mathbf{y}_t)$ through the importance weights:
   $w_t^{(i)} \propto \frac{p_{N_z}(\mathbf{y}_t|\mathbf{X}_{0:t}^{(i)}, \mathbf{y}_{0:t-1})\, p_{N_z}(\mathbf{X}_t^{(i)}|\mathbf{X}_{0:t-1}^{(i)}, \mathbf{y}_{0:t-1})}{\pi(\mathbf{X}_t^{(i)}|\mathbf{X}_{0:t-1}^{(i)}, \mathbf{y}_{0:t-1})}.$
2. Resample $\{\mathbf{X}_t^{(i)}, \mathbf{Z}_t^{(i,j)}\}$ according to $w_t^{(i)}$.
3. Approximation of $p(\mathbf{Z}_t|\mathbf{X}_{0:t}, \mathbf{y}_{0:t})$ through the importance weights:
   $\bar{q}_t^{(i,j)} \propto p(\mathbf{y}_t|\mathbf{X}_t^{(i)}, \mathbf{Z}_t^{(i,j)}).$
4. Generate $\mathbf{X}_{t+1}^{(i)}$ according to $\pi(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}^{(i)}, \mathbf{y}_{0:t})$.
5. Approximation of $p(\mathbf{Z}_t|\mathbf{X}_{0:t+1}, \mathbf{y}_{0:t})$ through the importance weights
   $q_t^{(i,j)} = \bar{q}_t^{(i,j)}\, p(\mathbf{X}_{t+1}^{(i)}|\mathbf{X}_t^{(i)}, \mathbf{Z}_t^{(i,j)}).$
6. Resample $\mathbf{Z}_t^{(i,j)}$ according to $q_t^{(i,j)}$.
7. Generation of particles $\mathbf{Z}_{t+1}^{(i,j)}$ according to the proposal
   $\pi(\mathbf{Z}_{t+1}|\mathbf{X}_{0:t+1}^{(i)}, \mathbf{Z}_t^{(i,j)}, \mathbf{y}_{0:t})$.
```

will be the joint state of the individuals, $\mathbf{Z}$ that of the groups. It follows that all the distributions introduced above have been designed and engineered to fit into the new context.

### 6.4.3 Joint Individual-Group Tracking

Let $\mathbf{X}_t = \{\mathbf{x}_t^k\}_{k=1}^K$ be the joint state of the $K$ individuals at time $t$ and $\mathbf{Z}_t = \{\mathbf{z}_t^k\}_{k=1}^K$ with $\mathbf{z}_t^k \in \{0, 1, \ldots, G\}$ be the the joint state of the $G$ groups ($K$ and $G$ may vary over time). We define $\mathbf{x}_t^k = (x_t, y_t, \dot{x}_t, \dot{y}_t)$ (individual positions and velocities) and $\mathbf{z}_t^k$ as the group's label for the $k$-th individual. As an example, suppose we have 5 individuals and 2 groups at time $t$: with $\mathbf{Z}_t = [1, 1, 2, 2, 0]^T$ we indicate that the first two individuals belong to the first group, the third and fourth individual are in the second group, and the fifth individual is a singleton. Compared with the group state space defined in Co-PF, the label-based state space is discrete and low-dimensional. In the example just presented, the group state space in Co-PF would be 30-dimensional in contrast with the current one that has 5 dimensions. The size of the whole state space for the DEEPER-JIGT is 25 dimension. This highlights the complexity of performing inference in such state space.

| Distribution | Analyt. | Approx. |
|---|---|---|
| $\pi(\mathbf{X}_{t+1}\|\mathbf{X}_{0:t}, \mathbf{y}_{0:t})$ | ✓ | ✓ |
| $p(\mathbf{y}_t\|\mathbf{X}_t, \mathbf{Z}_t)$ | ✓ | ✗ |
| $p(\mathbf{X}_{t+1}\|\mathbf{X}_t, \mathbf{Z}_t)$ | ✓ | ✗ |
| $\pi(\mathbf{Z}_{t+1}\|\mathbf{X}_{0:t+1}, \mathbf{Z}_t, \mathbf{y}_{0:t})$ | ✗ | ✓ |

**Table 6.2.** Probability $p(\cdot)$ and proposal $\pi(\cdot)$ distributions that have to be designed in the DPF. The second and third columns identify which distributions are evaluated and sampled, respectively.

The customization of the DPF algorithm for our tracking scenario requires an appropriate design of the probability distributions highlighted in gray in Algorithm 4. As usual in the particle filtering strategies, distributions may have an analytical form, and/or they can be approximated by particles' sets. In general, one prefers the latter case as this allows one to deal with arbitrarily complex distributions. Analytical functions are usually simpler, typically with Gaussian profiles, but this reduces the expressiveness of the tracking posterior. For more complex analytic functions, exact inference is intractable, due to the marginalization over the state space of the filtering formulation. From Algorithm 4, we can notice which distribution has to be represented in analytical form and which one is sampled. Table 6.2 summarized that distinction. Each distributions will be defined in the following.

**Individual Proposal $\pi(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{0:t})$.**

This distribution models the dynamics of the individuals in the same way we have already seen in Section 2.2.2. Inspired by [178], we adopt the notion of *composite* proposal, incorporating two sources of information:

$$\pi(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{0:t+1}) = \alpha\, \pi(\mathbf{X}_{t+1}|\mathbf{X}_t) + (1-\alpha)\, \pi_{\mathrm{det}}(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{t+1}).$$

Here, the first part assumes Markovianity between $\mathbf{X}$'s and conditional independence with the observation $\mathbf{y}_t$, and adopts a locally linear dynamics with Gaussian noise:

$$\mathbf{x}_{t+1}^k = A\mathbf{x}_t^k + \eta \quad \text{with } A = \begin{bmatrix} 1 & 0 & T & 0 \\ 0 & 1 & 0 & T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where $T$ is the sampling interval and $\eta \sim \mathcal{N}(\mathbf{0}, \Sigma^k)$ is the noise. Therefore, $\mathbf{x}_{t+1}^k \sim \mathcal{N}(A\mathbf{x}_t^k, \Sigma^k)$, that is easy to evaluate and sample from. Assuming independence between individuals motion, we have:

$$\pi(\mathbf{X}_{t+1}|\mathbf{X}_t) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_{t+1}^k|A\mathbf{x}_t^k, \Sigma^k) \tag{6.14}$$

that is, a multivariate Gaussian distribution with block-diagonal covariance matrix: $\mathrm{diag}(\Sigma^1, \Sigma^2, \ldots, \Sigma^K)$. We can assume $\Sigma^k = \Sigma$ for each $k = 1, \ldots, K$, supposing individuals having usually similar motions. Note that we can also incorporate the dynamical model proposed in the HJS filter (Section 2.2.1), that considers the exclusion principle between individuals.

The second part $\pi_{\mathrm{det}}(\mathbf{X}_{t+1}|\mathbf{X}_{0:t}, \mathbf{y}_{t+1})$ presumes the presence of a detector[1]. This distribution is defined as a multivariate Gaussian distribution with the same covariance matrix of Eq. 6.14 and the positions of the detections associated to each target as means. The parameter $\alpha$ is decided once and kept fixed for all the experiments.

### Joint Observation Distribution $p(\mathbf{y}_t|\mathbf{X}_t, \mathbf{Z}_t)$.

We adopt a standard template-based technique [46], where the goal is to find the hypothesis that is most similar to a template of the object that is being tracked. To make standard observation models suitable for our framework, we re-write the joint observation distribution as follows:

$$p(\mathbf{y}_t|\mathbf{X}_t, \mathbf{Z}_t) \propto p(\mathbf{Z}_t|\mathbf{y}_t, \mathbf{X}_t)\, p(\mathbf{y}_t|\mathbf{X}_t). \tag{6.15}$$

In this way, we can model $p(\mathbf{y}_t|\mathbf{X}_t)$ as in standard particle filtering approaches that we have seen in the previous chapters. For simplicity, we define it assuming independence between targets as follows:

$$p(\mathbf{y}_t|\mathbf{X}_t) = \prod_{k=1}^{K} p(\mathbf{y}_t|\mathbf{x}_t^k) \propto \prod_{k=1}^{K} exp(-\lambda_{d_y}\, d_y(f(\mathbf{y}_t, \mathbf{x}_t^k), \tau^k))$$

where $d_y$ is a distance between features, $f(\mathbf{y}_t, \mathbf{x}_t^k)$ extract features from the current bounding box in the image given by $\mathbf{x}_t^k$ and $\tau^k$ is the template of the $k$-th individual. In our experiments, we use the Bhattacharyya distance between RGB color histograms and the template is never updated. Note that plenty of more sophisticated techniques fitting our framework are available in the literature.

We also assume conditional independence between $\mathbf{Z}_t$ and $\mathbf{y}_t$, *i.e.*, $p(\mathbf{Z}_t|\mathbf{y}_t, \mathbf{X}_t, ) = p(\mathbf{Z}_t|\mathbf{X}_t)$. This term models the likelihood that $\mathbf{Z}_t$ has been generated from $\mathbf{X}_t$. In terms of particles, each group hypothesis $\mathbf{Z}_t^{(i,j)}$ can be seen as a clustering hypothesis of the data $\mathbf{X}_t^{(i)}$. Hence, $p(\mathbf{Z}_t|\mathbf{X}_t)$ can be formulated in terms of *cluster validity* evaluation as follows:

$$p(\mathbf{Z}_t|\mathbf{X}_t) \propto exp(-\lambda_{d_{cl}}\, d_{cl}(\mathbf{Z}_t, \mathbf{X}_t))$$

where $d_{cl}(\mathbf{Z}_t, \mathbf{X}_t)$ is a cluster validity measurement of the hypothesis $\mathbf{Z}_t$ with respect to $\mathbf{X}_t$. Among the different cluster validity measurements, we choose the Davies-Bouldin index [66], because of its simplicity and versatility, but this does not exclude the use of other cluster validity indexes.

[1] In our experiments, we used as detections the perturbed ground truth by adding false positives and false negatives.

**Joint Individual Distribution** $p(\mathbf{X}_{t+1}|\mathbf{X}_t, \mathbf{Z}_t)$**.**

This distribution models the dynamics of the individual taking into account the presence of the group:

$$\mathbf{x}_{t+1}^k = \mathbf{x}_t^k + B\mathbf{g}_t^k + \eta \tag{6.16}$$

where

$$B = \begin{bmatrix} 0 & 0 & T & 0 \\ 0 & 0 & 0 & T \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{g}_t^k = \frac{\sum_{l=1}^K \mathbf{x}_t^l \, \mathbb{I}(\mathbf{z}_t^k == \mathbf{z}_t^l)}{\sum_{l=1}^K \mathbb{I}(\mathbf{z}_t^k == \mathbf{z}_t^l)}$$

$\mathbb{I}(\cdot)$ is the indicator function and $\mathbf{g}_t^k$ is the position and velocity of the group the $k$-th individual belongs to. This term mirrors the fact that individuals in the same group should have similar dynamics. Notice that the matrix $B$ selects only the velocity components of the vector $\mathbf{g}_t^k$. Thus, the groups position does not affect the individuals position.

Similarly to Eq. 6.14, the resulting probability distribution is:

$$p(\mathbf{X}_{t+1}|\mathbf{X}_t, \mathbf{Z}_t) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}_{t+1}^k|\mathbf{x}_t^k + B\mathbf{g}_t^k, \Sigma)$$

that is again a multivariate Gaussian distribution with block-diagonal covariance matrix.

**Joint Group Proposal** $\pi(\mathbf{Z}_{t+1}|\mathbf{X}_{0:t+1}, \mathbf{Z}_t, \mathbf{y}_{0:t})$**.**

The joint group proposal models the dynamics of the groups and how samples are generated in the group state space. It is very important do define it in a smart way, in order to avoid an exhaustive exploration of the combinatorial state space. This point will be clearer in the experiments. We define the joint group proposal models as follows:

$$\pi(\mathbf{Z}_{t+1}|\mathbf{X}_{0:t+1}, \mathbf{Z}_t, \mathbf{y}_{0:t}) = f(\prod_{g=1}^G \pi(e_{t+1}^g|\mathbf{X}_{0:t+1}, \mathbf{Z}_t, \mathbf{y}_{0:t}), \mathbf{Z}_t) \tag{6.17}$$

$$= f(\prod_{g=1}^G \pi(e_{t+1}^g|\mathbf{X}_{0:t+1}, g_t, g_t', \mathbf{y}_{0:t}), \mathbf{Z}_t) \tag{6.18}$$

where the *surrogate* distribution $\pi(e_{t+1}^g|\mathbf{X}_{0:t+1}, \mathbf{Z}_t, \mathbf{y}_{0:t})$ in Eq. 6.17 operates by assigning probabilities on the *events* related to the $g$-th group, *i.e.*, $e^g \in \{\text{Merge}, \text{Split}, \text{None}\}$. In other words, given a group configuration $\mathbf{Z}_t$, whose individuals moved as recorded in $\mathbf{X}_{0:t+1}$, we want to model the probability that a merge or split event does occur, or that the group assignment of each individual remains unchanged. To simplify the modeling, the surrogate is rewritten as in Eq. 6.18, considering only interactions between a group $g$ and its nearest group $g'$. The deterministic function $f$ translates a selected event in a novel configuration

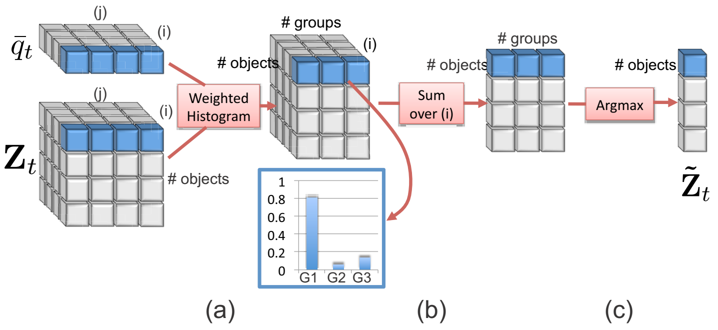


**Fig. 6.6.** State $\tilde{\mathbf{Z}}_t$ estimate that deals with discrete labels

$\mathbf{Z}_{t+1}$, changing the label assignment of $\mathbf{Z}_t$, enlarging or diminishing its size if novel objects (dis)appear. Note that in our approach, a group is an entity formed at least by two individuals.

The distribution $\pi(e^g_{t+1}|\mathbf{X}_{0:t+1}, g_t, g'_t, \mathbf{y}_{0:t})$ is offline learned, adopting the multinomial logistic regression. To this end, a set of possible scenarios containing events have been simulated and labelled. We use as features the inter-group distance between $g$ and the nearest group $g'$, considering their positions and extensions ($d_{KL}$, Kullback-Leibler distance between Gaussians) and velocities ($d_v$, Euclidean distance), and the intra-group variance between the positions of the individuals in the $g$-th group ($d_{\text{intra}}$). Thus, the input of the multinomial logistic regression is a 6-dimensional vector, *i.e.*, $(d_{KL}, d_v, d_{\text{intra}})$ for time $t$ and $t+1$.

Once the model has been trained, performing inference is straightforward. Given an existing group $g$, $(d_{KL}, d_v, d_{\text{intra}})$ for time $t$ and $t+1$ are computed and fed into the classifier, obtaining an estimate of $\pi(e^g_{t+1}|\mathbf{X}_{0:t+1}, g_t, {g'}_t, \mathbf{y}_{0:t})$. A new event $e^g_{t+1}$ is sampled from that distribution. Note that sampling from it is easy and efficient, because it is discrete and the set of possible events is relatively small. Once the event $e^g_{t+1}$ has been sampled from the proposal distribution, the function $f(\cdot)$ performs the selected event to generate $\mathbf{Z}_{t+1}$.

In addition, we add a prior over the events in order to reduce the merge between too-large groups. The prior is defined as $\mathcal{N}(|\mathbf{e}^g_{t+1}|; \mu, \sigma)$ where $|\mathbf{e}^g_{t+1}|$ is the size of the $g$-th group after the event $g$ (in the experiments, $\mu = 1$ and $\sigma = 1.5$).

**State Estimate.** In this section, we describe how to estimate the most likely joint state. The joint probability distribution $p(\mathbf{Z}_t, \mathbf{X}_{0:t}|\mathbf{y}_{0:t})$ can be estimated by the DPF once defined each probability distribution in Table 6.2. The joint state is usually defined as the expected value of the state under a certain distribution, that is, $\widetilde{\mathbf{X}}_t = \mathbb{E}_{p(\mathbf{X}_t|\mathbf{y}_{0:t})}[\mathbf{X}_t]$ and $\widetilde{\mathbf{Z}}_t = \mathbb{E}_{p(\mathbf{Z}_t|\mathbf{y}_{0:t})}[\mathbf{Z}_t]$. Using the empirical approximation given by the DPF, we can easily estimate $\mathbf{X}_t$ as $\tilde{\mathbf{X}}_t = \sum_{i=1}^{N_x} w_t^{(i)} \mathbf{X}_t^{(i)}$.

Since the domain $\mathbf{Z}_t$ is based on discrete labels, the expectation operation cannot be performed directly. Instead, we compute a distribution over the possible labels as depicted in Figure 6.6. Starting from the matrices $\mathbf{Z}_t$ and $\bar{q}_t$, we compute the following distribution for the $k$-th individual as weighted histogram:

$$\mathrm{Wh}^{k,(i,g)} = \sum_{j=1}^{N_z} \bar{q}_t^{(i,j)} \mathbb{I}(\mathbf{z}_t^{k,(i,j)} == g).$$

This gives a similar representation of the sum over $j$ but it considers labels $g$ (step (a) in Figure 6.6). Then, each $\mathrm{Wh}^{k,(i,g)}$ is summed over $i$ (step (b) in Figure 6.6), and we take the maximum likelihood estimate of the association between groups and individuals to obtain $\tilde{\mathbf{Z}}_t$ (step (c) in Figure 6.6).

## 6.5 Experiments

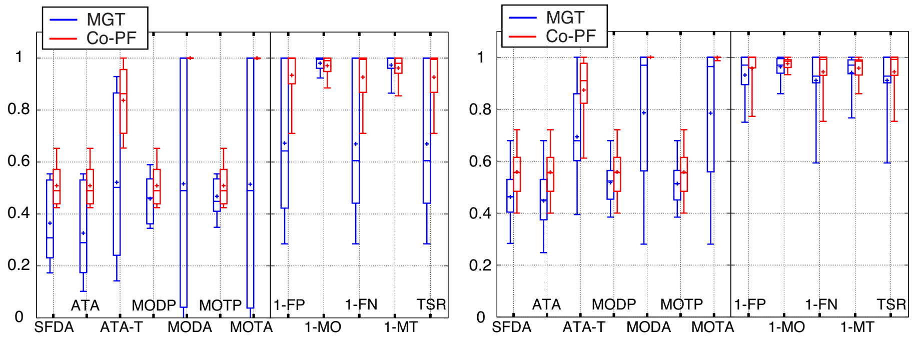
In this section, Co-PF and DEEPER-JIGT have been evaluated in terms of quantitative and qualitative results. Unfortunately, a fair comparison of these two models is not possible, because groups in Co-PF are manually initialized, while DEEPER-JIGT makes up groups from the scratch. In addition, DEEPER-JIGT is able to handle group events while Co-PF cannot. Thus, a comparison using the Friends Meet dataset will carry out expected results, that is, DEEPER-JIGT would be better. Instead, if we had compared them in the PETS sequence used for Co-PF, this would favorite Co-PF because in the selected sequences groups do not neither split nor merge.

### 6.5.1 CoPF

The approach has been evaluated on synthetic data and two publicly available datasets: PETS 2006 [4] and PETS 2009 [6]. We carried out a comparative analysis with respect to the group tracking without the proposed collaboration stage (called here MGT), highlighting that Co-PF is better to deal with intra- and inter-group occlusion. Other approaches have not been taken into account because of the lack of: 1) on-line available code for any of the approaches in the state of the art 2) a shared, labelled, dataset.

The simulations on the synthetic test set are carried out, in order to build statistics on ground truthed sequences. The test set is built to emulate the scenarios in the PETS datasets by using the same background and the same calibration data. Each sequence contains static images of people "walking" in the environment and forming groups. We artificially create a set of 26 sequences (13 for each dataset), choosing two different points of view in order to deal with variably scaled people: the first camera is closed to the people, while the second one is far. The number of people and the number of groups vary in different sequences from 3 to 20 and from 1 to 5, respectively. The number of person in a group varies from 2 to 6. The parameters are set as follows: $\sigma_\mu = 0.05$, $\sigma_\lambda = 0.05$, $\sigma_\theta = \pi/40$, 256 bin are used for the HSV histogram, $\alpha_1 = \alpha_2 = 0.5$, $\tau_b = 0.5$, $\tau_a = 2.5$.

A comparison has been done between the Co-PF with $N = 50$ and $N_g = 50$ (the number of particles for each group) and MGT with $N'_g \approx N_g + N \cdot \frac{K^2}{G^2 \cdot C}$, where $C = 5$ has been empirically chosen, $K$ and $G$ are the number of people and groups, respectively. In this way, the computational burden of the two methods is similar. To evaluate the performance on the synthetic test set, we adopt the same



(a) PETS 2006 results

(b) PETS 2009 results

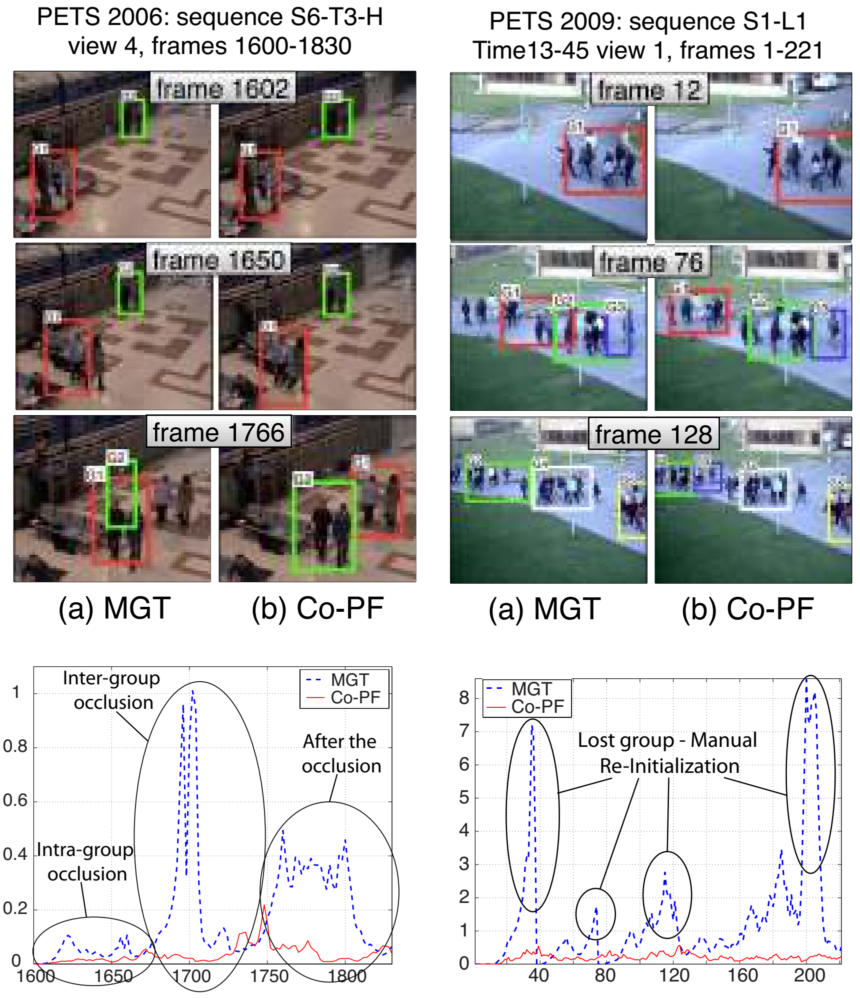
**Fig. 6.7.** Statistics on the synthetic test set.

measurements discussed in Section 2.4.1: False Positive (FP), Multiple Objects (MO), False Negative (FN), and Tracking Success Rate (TSR). Additionally, we use other standard metrics presented in [126] using the code proposed by the authors: Average Tracking Accuracy (ATA), ATA Thresholded (ATA-T), Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), Multiple Object Detection Accuracy (MODA), Sequence Frame Detection Accuracy (SFDA), and Multiple Object Detection Precision (MODP). Note that the detection measurements (MODP, MODA, SFDA) consider the track estimates as group detections, without taking in account the identifiers and thus the temporal links.

For each measure, a *boxplot* representation is given [126], where the box is defined by the 1st quartile, median, and the 3rd quartile; the extremities outside the box are the smallest and largest value, and the ”+” is the mean value. The comparison (Figure 6.7(a) and Figure 6.7(b)) shows that in the PETS2006 synthetic dataset our Co-PF strongly outperforms the MGT in terms of all the measures. Even though the PETS2009 sequences are slightly harder, Co-PF often succeeds where MGT fails, yielding to higher performances.

Moreover, we perform the test on portions of the PETS datasets, using the same settings. We consider sequences where the groups were not subjected to splits or merges, in order to stress the capability of tracking group entities with intra- and inter-group occlusions. Initialization of groups has been done by fitting the $\mu^g$ and $\Sigma^g$ to the projections of the individuals new entries on the ground plane. If lost, a group is manually reinitialized. We show here two representative examples. In real scenarios, MGT is not able to deal completely with the intra- and inter-group dynamics (Figure 6.8(a)). On the other hand, Co-PF exploits the MOT results, enriching the posterior knowledge given by the MGT (Figure 6.8(b)).

To give further support to our Co-PF, we evaluate the uncertainty of the particle filters defined as the variance of the importance weights [161]. Figure 6.8(c) depicts that the MGT uncertainty is peaked when an intra- and inter-group occlusion occurs. After the occlusion the uncertainty is high because the track is erroneously lost (two tracks on a single group). Figure 6.8(d) shows a similar behavior of Figure 6.8(c), highlighting that the MGT looses the tracks several times.



(c) PETS 2006

(d) PETS 2009

**Fig. 6.8.** Comparison of MGT (first and third column) and Co-PF (second and fourth column) on PETS 2006 and PETS 2009. The second row compares the PF uncertainty [161] in the two experiments.

### 6.5.2 DEEPER-JIGT

This section shows the potentialities of DEEPER-JIGT in performing joint individual-group tracking on different datasets, while investigating the effects of the mutual support of the group and the individual tracking processes. The structured filtering architecture of DEEPER-JIGT allowed to achieve this goal, by inhibiting conditional dependencies in distributions where mixed terms (**X** and **Z**) do appear.

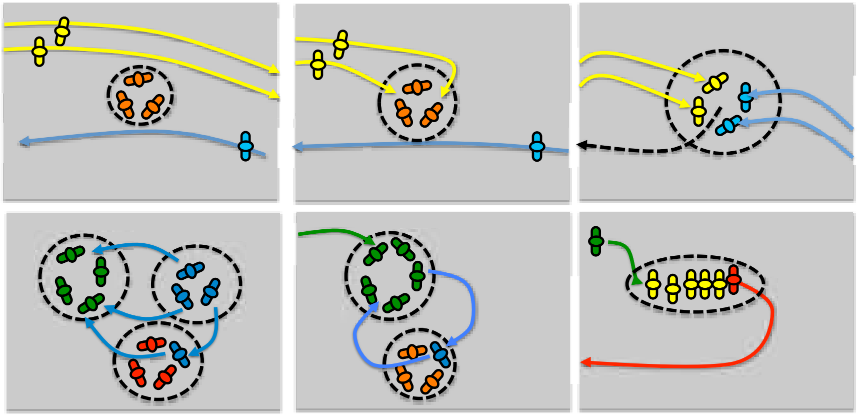


**Fig. 6.9.** Typical scenarios in the *Friends Meet* dataset: merge and split between groups, queue, and complex situations.

**Datasets.** The ideal benchmark should handle a scenario where *labelled* groups of people are evolving, appearing and disappearing spontaneously, experiencing split and merge events. This correspond to cocktail party-like situations, *i.e.*, focusing on social areas where people arrive alone or with other people, move from one group to another, stay still while conversing, etc. Nowadays, such a picture is missing, since almost all the existent datasets with labelled groups report different situations, mainly wandering people following a main flow direction (*e.g.*, [190]). In this case, groups are mostly limited to very few people (mostly couples) and the frequency of merge and split is low.

For these reasons, we propose a novel dataset, freely downloadable at `http://goo.gl/cFXCG`, dubbed *Friends Meet*. It is composed by 53 sequences, for a total of 16286 frames. The sequences are partitioned in a synthetic set (28 sequences, 200 frames each), with the aim of stressing tracking strategies in capturing group events, without any complex object representation (simple colored blobs), and a real dataset. In the *synthetic* set, 18 sequences are simple, containing 1-2 events with 4-10 individuals; the other 10 sequences are more challenging, with 10-16 individuals involved in multiple events.

The *real* set focuses on an outdoor area where people usually meet during coffee breaks. This area has been recorded and annotated by an expert for one month. The expert reported the events appeared more frequently, building a screenplay where these events are summarized in order to limit the dataset size. Therefore, the screenplay was played by students and employees, resulting in 15 sequences of different length (between 30 sec. to 1.5 minutes), judged by the expert as sufficiently realistic. In total, the sequences contain from 3 to 11 individuals, and all of them are ground truthed with individual and group information. Some typical scenes are depicted in Fig. 6.9.

In addition, we provide both quantitative and qualitative results using the BIWI dataset [190], even though it is not well-suited for our method because group events are absent.

**Evaluation Metrics.** The evaluation of DEEPER-JIGT considered the individual tracking and the group tracking results. Unfortunately, while there exists a lot of standard evaluation metrics for individual tracking, there are no widely-accepted measures for group tracking. For this reason, we customize the metrics proposed in [38, 228] to deal with groups.

For individual tracking, we employ Mean Square Error (MSE) over the positions of the individuals, and its standard deviation. The group results have been evaluated by adapting the metrics proposed in [228] for detection (False Positive (FP) and False Negative (FN)) and in [38] for tracking (Multi-Object Tracking Precision (MOTP) and Accuracy (MOTA)). Those tracking metrics relie on the definition of person bounding box (rectangular region) and intersection between the estimate and the ground truth bounding boxes to determine if we have a true positive, a false positive or false negative. When going from individuals to groups the concept of buonding box is replaced by a more complex structure, that can be a ellipse or a convex hull. We propose to sobstitute the notion of person bounding box to that of convex hull around the members of groups. Each memeber of the group (that has its own extent) contributes in the creation of the convex hull. Therefore, given the convex hull of the estimated group and of the ground truth group, the intersection operations among bounding boxes translate naturally in that of convex hulls. If the estimated convex hull overlaps with the ground truth convex hull, we have a true positive if the overlapping threshold is fulfilled. We have a false negative if the overlapping threshold is not fulfilled and any other estimated convex hulls fulfill the overlapping threshold. We have a false positive if the ground truth group already have an associated estimate and another estimate group tries to detect it.

We also introduced the Group Detection Success Rate (GDSR) as the detection rate over time of the correctly detected groups. A group is *correct* if at least the 60% of its members are detected [62]. For example, if there are 2 groups and 2 singletons in the scene, then the ground truth state is $\mathbf{Z}_t = [1, 1, 1, 1, 2, 2, 0, 0]$. Assuming that the state estimate is $\tilde{\mathbf{Z}}_t = [1, 1, 0, 0, 3, 3, 0, 0]$, we will match the group 1 of 4 people in $\mathbf{Z}_t$ with the group 1 of 2 people in $\tilde{\mathbf{Z}}_t$ and he group 2 of 2 people in $\mathbf{Z}_t$ with the group 3 of 2 people in $\tilde{\mathbf{Z}}_t$. The first matched group will be not correctly detected, because the system detected 50% of its members, while the second matched group will be correct. In this case, the GDSR for the current time step will be equal to 1/2 (averaged over the number of groups in $\mathbf{Z}_t$). After computing it for every frame, the GDSR is averaged over the number of time steps.

**Results.** The evaluation focused first on the synthetic part of the Friends Meet dataset. For the investigation of the mutual support of the group and the individual tracking processes, we build three variants of DEEPER-JIGT, that is, `VAR1`, `VAR2` and `VAR3` .

`VAR1` assumes $p(\mathbf{X}_{t+1}|\mathbf{X}_t, \mathbf{Z}_t) = p(\mathbf{X}_{t+1}|\mathbf{X}_t)$, inhibiting the contribute of the group in defining the dynamics of the individual, by canceling out the $B\mathbf{g}_t^k$ term of Eq. 6.16. `VAR2` is equal to `VAR1`, assuming in addition $\pi(\mathbf{Z}_{t+1}|\mathbf{X}_{0:t+1}, \mathbf{Z}_t, \mathbf{y}_{0:t}) = \pi(\mathbf{Z}_{t+1}|\mathbf{Z}_t, \mathbf{y}_{0:t})$, that is, suppressing the knowledge of the individual state in promoting events for the group evolution. In practice, instead of sampling from the surrogate distribution of events, we sampled from the combinatorial space of possible configurations of the group hypothesis, supposing them distributed in a uniform fashion. From DEEPER-JIGT to VAR2, we can notice that the distributions become conditionally independent, and thus sampling is performed indipendently in each state space. Only the observation model links them. Finally, VAR3 is VAR2 with $p(\mathbf{y}_t|\mathbf{X}_t, \mathbf{Z}_t) = p(\mathbf{y}_t|\mathbf{X}_t)$, blocking the contribution of the clustering evaluation, i.e., fixing $p(\mathbf{Z}_t|\mathbf{y}_t, \mathbf{X}_t) = 1$ in Eq. 6.15. This way, the model judges the individuals, but not how well they fit in groups hypotheses. In practice, this variant separates the individual tracking from the group tracking in two different particle filters.

| | MSE px<br>(std) | 1-FP | 1-FN | GDSR | MOTP px | MOTA |
|---|---|---|---|---|---|---|
| DEEPER-JIGT | **2.18**<br>(**4.96**) | 93.74% | **82.94**% | **79.65**% | **16.66** | **57.28**% |
| VAR1 | 2.19<br>(5.46) | **93.77**% | 81.86% | 78.25% | 17.74 | 55.99% |
| VAR2 | 3.72<br>(11.81) | 82.11% | 51.61% | 48.09% | 151.03 | 33.53% |
| VAR3 | 2.52<br>(8.35) | 65.09% | 24.56% | 18.89% | 397.85 | 4.86% |

**Table 6.3.** Results on synthetic sequences: individual tracking (column 2), group detection (columns 3-5) and group tracking (column 6-7). For MSE and MOTP (in pixels), the lower the better.

The results on the synthetic data are summarized in Table 6.3. The first significant message is clear: the components of DEEPER-JIGT are all needed for reaching the best performances. Moreover, all the performance measures decrease when incrementally pruning away connections between the individuals and the groups (from VAR1 to VAR3).

Actually, the performance of VAR1 tells that in a joint individual-group tracking framework, the individual dynamics should consider the influence that the group exerts on the single person. This helps just a little the group description, while it is uninfluential if we focus on the individual tracking only. We think that this relationship could be exploited more effectively if more advanced group-driven dynamics are injected, *e.g.*, [62, 190, 270].

The performance of VAR2 suggests that the dynamics of a group (intended as the possibility of splitting or merging) cannot be treated as an independent process, and must necessarily be linked to the behavior of the single individuals. This is intuitive, and is beneficial for both individual and group tracking. Even in this case, social grouping mechanisms [62, 190, 270] can boost the performances. The performance of VAR3 is the most enlightening: it shows that for modeling groups and individuals a joint treatment is highly recommendable, being the performances of the two separate processes strongly inferior to DEEPER-JIGT.

The second analysis takes into account the real datasets. Since these datasets are very challenging for tracking, due to occlusions and low resolution, a track is re-initialized from the ground truth when the target is lost (distance of 0.6 meters). The mean re-initialization rate for a target is 3.2% for the real FM dataset.

We compare DEEPER-JIGT against VAR3 and DEEPER-JIGT.2. In DEEPER-JIGT.2, we assume that $p(\mathbf{X}_{0:t}|\mathbf{y}_t)$ is completely known, that is, at each time the individual tracker is initialized from the ground truth. In other words, this variant of the algorithm evaluates the method when very low uncertainty on individual tracking is present, thus representing an upper bound on the group performances.

a) FM dataset

| | 1-FP | 1-FN | GDSR | MOTP m | MOTA |
|---|---|---|---|---|---|
| DEEPER-JIGT.2 | **97.05%** | **93.82%** | **88.46%** | **0.64** | **71.70%** |
| DEEPER-JIGT | 95.61% | 91.13% | 86.11% | 0.80 | 67.58% |
| VAR3 | 74.77% | 37.72% | 25.92% | 2.80 | 2.73% |

b) BIWI dataset

| | 1-FP | 1-FN | GDSR | MOTP m | MOTA |
|---|---|---|---|---|---|
| DEEPER-JIGT | 53.77% | **78.00%** | **53.59%** | **0.44** | **29.43%** |
| VAR3 | **60.55%** | 51.57% | 29.60% | 1.03 | 9.58% |

**Table 6.4.** Group results on a) the FM dataset and b) the BIWI dataset: group detection (columns 2-4) and group tracking (column 5-6). For MOTP (in meters), the lower the better.

The group tracking accuracies on the FM dataset and the BIWI dataset are summarized in Table 6.4(a-b). The table highlights the increase of the performance from VAR3 to DEEPER-JIGT. Differently from the synthetic scenario, the false positive rates (1−FP) of the different methods are close (Table 6.4(a)). The low value of 1−FN for VAR3 mirrors the fact that the method looses the 56% of the groups. The other metrics follow the trend of the results on the synthetic dataset.

Comparing DEEPER-JIGT and DEEPER-JIGT.2 (Table 6.4(a)), it is interesting to note that if $p(\mathbf{X}_{0:t}|\mathbf{y}_t)$ is known, we obtain very similar results. This means that the uncertainty in the process does not affect very much the joint individual-group tracking. Moreover, Table 6.4(b) shows that even if the BIWI dataset is harder due to the low resolution and does not contain groups event, DEEPER-JIGT is still able to get reasonable results.

Qualitative results of DEEPER-JIGT on FM dataset (rows 1-3) and BIWI dataset (row 4) are reported in Fig. 6.10 and the video at http://youtu.be/J_HDJflQATo. The figure shows different examples of merge (row 1), initialization and split (row 2), and more complex scene where multiple events occur (row 3). In the sequence seq_eth of BIWI dataset (row 4), we noticed that DEEPER-JIGT is able to capture groups of wandering people, even in the case of crowd.

## 6.6 Conclusions

In this chapter, the individual-group tracking problem has been presented and discussed extensively. Two solutions have been proposed to deal with the problem: Co-PF and DEEPER-JIGT.

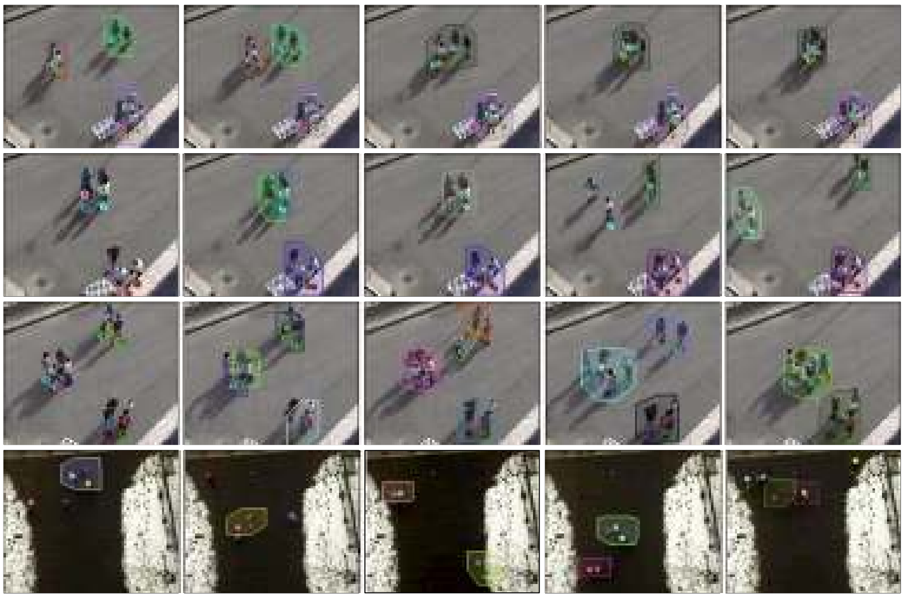


**Fig. 6.10.** Qualitative results on selected sequence of FM dataset (first three rows) and BIWI dataset (last row). 1st row: merging between two groups. 2nd row: split event, showing that when people is too far, they are detected as separate persons. 3rd row: a person move from one group to another, then a merge between two groups occurs.

In the Co-PF, two processes are involved: the group process simplifies groups as atomic entities, dealing successfully with occluding, multiple groups. The individual process captures how individuals move, refining the estimations done by the group process directly in the posterior distribution. In this way, the group tracking process can evolve in a more robust way.

After a close analysis of the Co-PF, we found out that it is too hard to embed the notion of groups events into it. Thus, we investigated an alternative, more promising direction, that is, the DEEPER-JIGT. This study promotes the joint online treatment of individuals and groups in tracking applications. Apart from sociological matters (people may decide to move differently whether they are alone or not), we showed here that this strategy is convenient *quantitatively*. As tracking strategy, we have been inspired from a brand-new filtering mechanism named Decentralized Particle Filter. The acronym DEEPER-JIGT mirrors its potentially deep customizability, that allows to tweak many filtering mechanisms, defining the dynamics of the group given the individual states and *vice-versa*, how observations are evaluated, etc., as modules of a serial framework. This is indeed a first attempt which proved to deserve further investigation. Next steps will be devoted to ameliorate the dynamics modules, possibly by embedding social force models as individual dynamics, improving how groups are evaluated, for example by importing social signal processing notions. At the same time, the inherent parallelization of the Decentralized Particle Filter, neglected here and ignored in the coding of DEEPER-JIGT, can be considered.

# Part IV

# Conclusions

# 7 Conclusions

The automatic analysis of images and videos is becoming every day more popular, given the recent success of several computer vision algorithms and systems. A lot of work has been done in the last decades in several applicative areas. This thesis is intended to cover several research aspects with application to video-surveillance and to propose some new interesting methods and algorithms to improve the state of the art. The main goal of this thesis is to cover a route[1] that tries to write an essay of what a researcher could find *beyond multi-target tracking*. In fact, the common denominators that characterize all the presented methods are two: 1) the use of the state-space used for tracking, that is the basic model for all the contributions and extensions proposed in this thesis. 2) The application to both video-surveillance and the analysis of social behaviors in videos. Let us retrace it from the beginning, making some *a-posteriori* questions and figuring out new possible directions and further research issues.

The travel started from (multi-)target tracking, whose aim is to localize object(s) in videos over space and time. Tracking has been extensively investigated in literature especially because we assisted to a growth of the number of CCTV cameras (and therefore data) in the world, but it still remain an open issue. Several methods have been proposed by the computer vision and signal processing communities. In this thesis, we kept the focus on few methods based on particle filtering, because of their advantages. We avoided a full description of the state-of-the-art algorithms and systems because it was out of the purpose of this thesis. Instead, we highlighted the specific problems that the research meets when facing multi-target tracking in computer vision, such as, filtering, data association, integration of the detections, appearance representation of the target, the template updating, and the occlusion handling. One of the most interesting problems we decided to handle is the latter. Our investigation went deeper in this direction, and the result was a novel mechanism of *online subjective feature selection* that can be embedded into any particle filter-based tracker. The simple idea was to distill a pool of features discriminating one individual with respect to the surrounding ones. This enables the tracker to deal with partial occlusions among targets. The results showed an increase in accuracy when embedding this strategy into a standard particle fil-

[1] We proposed to explore this route that was surely worth to investigate, but it is not unique.

ter. Its advantage is that any particle filter (not only hybrid joint-separable filter, which was the chosen testbed) that works on the joint state space can exploit the proposed method.

Many open issues still remain unsolved in target tracking. We are aware that a lot of work has to be done in future to have a complete system that is able to perform tracking on real scenarios with very little supervision by the human operator. For example, data association has not been well-investigated in this thesis. However, in the literature there exist a high number of techniques that can be used in combination with the standard trackers. Another interesting issue is the appearance description of the target. We partially investigated this issue, using one of the proposed descriptor for re-identification as appearance model for tracking. However, other issues still exist such as, how to online learn the model in order to avoid the drift of the tracker. Another interesting problem is how to deal with complete occlusions. One solution to this problem could be to perform person re-acquisition using the person re-identification methods proposed here.

Another interesting question raised in this thesis was about the path orthogonal to the standard target tracking solutions. In fact, the second stop of our journey was a bio-inspired model that takes inspiration from neuroscience theories of the human perception system. We proposed a decision-theoretic probabilistic graphical model for joint recognition, tracking and planning from gaze data. An attentional strategy is online learned to choose fixation points which lead to low uncertainty in the location of the target object. The model is composed of two interacting pathways that reflect the separation of information proposed by several works in the computational neuroscience literature. The identity pathway, responsible for comparing observations of the scene to an object template, consists of a 2-hidden layer deep network. The control pathway, responsible for aligning the object template with the full scene, is composed by a localization module and a fixation module.

Among several characteristics of the proposed model, its complementary view of the classic engineered computer vision approaches is one of the most interesting. The engineered computer vision approaches aim at solving specific problems to reach the state-of-the-art results. Instead, the proposed model takes inspiration from how the human perceptual system works, trying also to understand how the brain works. It can be seen also as a step towards a system that is able to think by itself from a more general perspective, instead of trying to deal with each problem separately. Several issues still remain open. For example, in our work we offline trained the appearance model. Existing particle filtering and stochastic optimization algorithms could be integrated in our model to train the 2-hidden layer deep network online. One of the most interesting avenues for future work is the construction of more abstract attentional strategies. In this work, we focused on attending regions of the visual field, but clearly one could attend to subsets of receptive fields or objects in the deep appearance model.

The attentional model was only a parenthesis of the perceptual point of view. The rest of the thesis followed the engineering direction. We investigated the extension of the multi-person tracking to the multi-camera setting, with particular interest in the non-overlapped camera views. The task, called person re-identification, has been analyzed in depth. The important contributions of this work were 1) the creation of a standard pipeline for re-identification, made by the following steps: images gathering, image selection, person segmentation, symmetry-based silhouette partition, descriptor extraction and accumulation, and signature matching. 2) The idea of gathering images over time to build a multi-shot descriptor made the method more robust. 3) The automatic subdivision in parts of the human silhouette enables us to perform part-based matching. We also found that among the steps of the pipeline, the descriptor extraction is very important. For this reason, we proposed three options called: Symmetry-Driven Accumulation of Local Features (SDLAF) , Histogram Plus Epitome (HPE) and Asymmetry-driven HPE (AHPE). SDALF is composed of three features: weighted color histogram, maximally stable color regions and recurrent highly-structured patches. HPE and AHPE are characterized by the use of the epitomic description that extract local motif by the human silhouette.

Each descriptor has its own advantages and drawbacks, that we have extensively examined in the experiments. The results we obtained on the public datasets are the state of the art, however they are still far from what we actually expect from a re-identification system. Therefore, future work is mandatory. One possible direction is to investigate different features and learning techniques to build the descriptor to obtain better results. However, the lack of a real, huge re-identification dataset makes the evaluation of the re-identification methods not down to earth. In fact, in the thesis we did test our descriptors on public datasets for comparison purposes, but we do not know what happen when dealing with a dataset with more than 1000 people. Thus, future work should be to gather a dataset that can be published and used by the community and test the current algorithms on that.

The last part of the route involved the extension of the standard target tracking, when the targets are both individuals and groups, first with group detection and then with individual-group tracking. We investigated how to detect social groups in a camera-monitored environment. We proposed a set of computer vision techniques for managing groups and group activities in a principled way, taking into account social psychology aspects (social signal processing) that define the human's acting. In this way, we moved from the un-personal objective point of view of the video camera capturing people as they were simple entities, to a new perspective where a subjective viewpoint of the individuals is taken into account. We showed how computer vision and social signal processing can collaborate for a new level of video surveillance research, also depicting the quality of the results such a collaboration can achieve. In this context, there are many future directions: for example, one can be interested in find finer type of relations between individuals inside a group (acquaintances, friends, engaged, *etc.*) using clustering techniques or infinite relational models. The subjective view frustum can be also used to monitor what are the most attended products in supermarkets, for data mining purposes.

In the context of social analysis of groups, the last chapter was devoted to the (joint) individual-group tracking problem. This thesis and the published works contained here are one of the very first attempt in literature to deal with this problem. Very few methods have been presented, especially by the computer vision community. Two complementary solutions have been proposed to deal with the problem, named Co-PF and DEEPER-JIGT. Co-PF handles the group and individual tracking separately and then the information is shared in a collaborative way when the individual estimate is available. This helps to keep low the computational burden, but as we will see later Co-PF cannot cope with groups events. On the other hand, DEEPER-JIGT is able to deal with group events, such as splitting, merging, initialization and deletion, in a probabilistic way. In contrast, most of the models proposed in literature usually model the group events by heuristic rules, yielding to a scarce generalization.

The results showed that individual-group tracking should be faced in the joint state space configuration, *i.e.*, using DEEPER-JIGT. The main reason is that Co-PF is only able to deal with atomic groups, and thus there are problem in case of split and merge. However, inference in the joint state space becomes very hard. DEEPER-JIGT is based on the decentralized particle filter that makes the problem more tractable, but yet not usable in practice. A lot of work has to be carried out before seeing an efficient version of joint individual-group tracking. Moreover, DEEPER-JIGT contains a delicate phase that enable the method to sample from the group proposal distribution. We used a supervised method to learn the distribution from which we sample, however we have found that in practice the training phase is not straightforward. Thus, future work could investigate how to embed a group proposal distribution that is trained in a unsupervised way. Furthermore, an online learning algorithm would even be more appreciated.

Summarizing, several improvements that are beyond multi-target tracking have been presented in this thesis. We did not claim that our solutions are the best possible ones, but we think that we have reached a good level of knowledge about the investigated problems and we are also aware of several future directions to improve the work. We hope that this thesis can be of inspiration to many researchers and engineers for further improvements and extensions, aiming at discovering more reliable solutions and models towards the next generation of automatic video surveillance systems.

# References

1. Caviar: Context aware vision using image-based active recognition dataset. http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/.
2. Ethz dataset for re-identification. http://www.liv.ic.unicamp.br/~wschwartz/datasets.html.
3. Gdet 2010: Group DETection dataset in a vending machines scenario. http://www.lorisbazzani.info/code-datasets/multi-camera-dataset/.
4. Pets 2006: Performance evaluation of tracking systems 2006 dataset. http://www.cvg.rdg.ac.uk/PETS2006.
5. Pets 2007: Performance evaluation of tracking systems 2007 dataset. http://www.pets2007.net.
6. Pets 2009: Performance evaluation of tracking systems 2009 dataset. http://www.pets2009.net.
7. UK home office, i-LIDS multiple camera tracking scenario definition. http://www.homeoffice.gov.uk/science-research/hosdb/i-lids/.
8. VIPeR dataset for re-identification. http://users.soe.ucsc.edu/~dgray/VIPeR.v1.0.zip.
9. Saad Ali and Mubarak Shah. Floor fields for tracking in high density crowd scenes. In *European Conference on Computer Vision*, pages 1–14, 2008.
10. M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1014 –1021, June 2009.
11. Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Monocular 3d pose estimation and tracking by detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 0:623–630, 2010.
12. M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Trans. on Signal Processing*, 50(2):174–188, 2002.
13. P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2001.
14. Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert Schapire. Gambling in a rigged casino: the adversarial multi-armed bandit problem. Technical Report NC2-TR-1998-025, 1998.
15. Alper Ayvaci, Michalis Raptis, and Stefano Soatto. Occlusion detection and motion estimation with convex optimization. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 100–108. 2010.

16. S. O. Ba and J. M. Odobez. A study on visual focus of attention recognition from head pose in a meeting room. In *MLMI*, pages 75–87, 2006.
17. B. Babenko, Ming-Hsuan Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1619 –1632, aug. 2011.
18. Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. Visual tracking with online multiple instance learning. In *Computer Vision and Pattern Recognition*, 2009.
19. Slawomir Bak, Etienne Corvee, Francois Bremond, and Monique Thonnat. Person Re-identification Using Haar-based and DCD-based Signature. In *2nd Workshop on Activity Monitoring by Multi-Camera Surveillance Systems, AMMCSS 2010*, Boston États-Unis, 08 2010.
20. Slawomir Bak, Etienne Corvee, Francois Bremond, and Monique Thonnat. Person Re-identification Using Spatial Covariance Regions of Human Body Parts. In *AVSS*, 08 2010.
21. Y. Bar-Shalom. *Tracking and data association*. 1987.
22. Y. Bar-Shalom, T. Formann, and M. Scheffe. Joint probability data association for multiple targets in clutter. *Proc. Conf. Information Science and Systems*, 1980.
23. Fracois Bardet, Thierry Chateau, and Datta Ramadasan. Illumination aware mcmc particle filter for long-term outdoor multi-object simultaneus tracking and classification. In *International Conference on Computer Vision*, 2009.
24. Luke Barrington, Tim K Marks, Janet Hui-Wen Hsiao, and Garrison W Cottrell. Nimble: a kernel density model of saccade-based visual memory. *Journal of Vision*, 8(14):1–14, 2008.
25. H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In *Proceedings of the European Conference on Computer Vision*, pages 404–417, 2006.
26. L. Bazzani. Caviar dataset for re-identification. `http://www.lorisbazzani.info/code-datasets/caviar4reid/`.
27. L. Bazzani, D. Bloisi, and V. Murino. A comparison of multi hypothesis kalman filter and particle filter for multi-target tracking. In *Performance Evaluation of Tracking and Surveillance workshop at CVPR*, pages 47–54, Miami, Florida, 2009.
28. L. Bazzani, M. Cristani, M. Bicego, and V. Murino. Online subjective feature selection for occlusion management in tracking applications. In *16th IEEE International Conference on Image Processing (ICIP)*, pages 3617 –3620, November 2009.
29. L. Bazzani, M. Cristani, and V. Murino. Collaborative particle filters for group tracking. In *17th IEEE International Conference on Image Processing (ICIP)*, pages 837 –840, September 2010.
30. L. Bazzani, M. Cristani, G. Pagetti, D. Tosato, G. Menegaz, and V. Murino. Analyzing groups: a social signaling perspective. In *Video Analytics for Business Intelligence*, Studies in Computational Intelligence. Springer-Verlag, 2012. in print.
31. L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by hpe signature. In *20th International Conference on Pattern Recognition (ICPR)*, pages 1413 –1416, August 2010. IBM Best Student Paper Award track: Computer Vision.
32. L. Bazzani, M. Cristani, A. Perina, and V. Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognition Letters*, 2011.
33. L. Bazzani, N. de Freitas, H. Larochelle, V. Murino, and J-A Ting. Learning attentional policies for object tracking and recognition in video with deep networks. In *International Conference on Machine Learning (ICML)*, 2011.
34. L. Bazzani, N. de Freitas, and J-A Ting. Learning attentional mechanisms for simultaneous object tracking and recognition with deep networks. In *Deep Learning and Unsupervised Feature Learning Workshop at NIPS*, Vancouver, Canada, 2010.

35. L. Bazzani, V. Murino, and M. Cristani. Decentralized particle filter for joint individual-group tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012. (to appear).
36. L. Bazzani, D. Tosato, M. Cristani, M. Farenzena, G. Pagetti, G. Menegaz, and V. Murino. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 2012. in print.
37. B. Benfold and I. Reid. Guiding visual surveillance by tracking human attention. In *Proceedings of the 20th British Machine Vision Conference*, September 2009.
38. Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *J. Image Video Process.*, 2008:1:1–1:10, January 2008.
39. N.D. Bird, O. Masoud, N.P. Papanikolopoulos, and A. Isaacs. Detection of loitering individuals in public transportation areas. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):167 – 177, 2005.
40. Michael J. Black and Allan D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26:63–84, January 1998.
41. S. S. Blackman. Multiple hypothesis tracking for multiple target tracking. *Aerospace and Electronic Systems Magazine, IEEE*, 19(1):5–18, 2004.
42. L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. Classification and Regression Trees. *Ann. Math. Statist.*, 19:293–325, 1984.
43. Michael D. Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *IEEE International Conference on Computer Vision (ICCV'09)*, October 2009. in press.
44. Michael D. Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1820–1833, 2011.
45. Eric Brochu, Vlad M. Cora, and Nando de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *CoRR*, abs/1012.2599, 2010.
46. L. G. Brown. A survey of image registration techniques. *ACM Comput. Surv.*, 24:325–376, December 1992.
47. M. Brown and D. G. Lowe. Unsupervised 3d object recognition and reconstruction in unordered datasets. In *Proceedings of the Fifth International Conference on 3-D Digital Imaging and Modeling*, pages 56–63, Washington, DC, USA, 2005. IEEE Computer Society.
48. M. A. Brubaker, L. Sigal, and D. J. Fleet. Video-based people tracking. *Handbook of Ambient Intelligence and Smart Environments*, pages 57–87, 2010.
49. Nicholas J. Butko and Javier R. Movellan. I-POMDP: An infomax model of eye movement. In *Proceedings of the International Conference on Development and Learning (ICDL 2008)*, August 2008.
50. Yizheng Cai, Nando de Freitas, and James J. Little. Robust visual tracking for multiple targets. In *European Conference on Computer Vision*, pages 107–118, 2006.
51. M. C. Chang, N. Krahnstoever, and W. Ge. Probabilistic group-level motion analysis and scenario recognition. In *IEEE International Conference on Computer Vision (ICCV)*, page 8, Nov. 2011.
52. Tianshi Chen, T. B. Schon, H. Ohlsson, and L. Ljung. Decentralized particle filter with arbitrary state decomposition. *Signal Processing, IEEE Transactions on*, 59(2):465 –478, February 2011.

53. D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conference (BMVC)*, 2011.
54. M. Cho and K. M. Lee. Bilateral symmetry detection and segmentation via symmetry-growing. In *British Machine Vision Conference*, 2009.
55. W. G. Choi, K. Shahid, and S. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. pages 1282–1289, 2009.
56. T. Choudhury and A. Pentland. The sociometer: A wearable device for understanding human networks. In *CSCW - Workshop on ACCUCE*, 2002.
57. D. Clark, I.T. Ruiz, Y. Petillot, and J. Bell. Particle phd filter multiple target tracking in sonar image. *Aerospace and Electronic Systems, IEEE Transactions on*, 43(1):409 –416, january 2007.
58. J. F. Cohn. Foundations of human computing: facial expression and emotion. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 233–238. ACM, 2006.
59. R. T. Collins, Y. Liu, and M. Leordeanu. Online selection of discriminative tracking features. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 27(10):1631–1643, 2005.
60. John Colombo. The development of visual attention in infancy. *Annual Review of Psychology*, pages 337–367, 2001.
61. Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003.
62. M. Cristani, L. Bazzani, G. Pagetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of f-formations. In *British Machine Vision Conference (BMVC)*, 2011.
63. M. Cristani, G. Pagetti, A. Vinciarelli, L. Bazzani, G. Menegaz, and V. Murino. Towards computational proxemics: Inferring social relations from interpersonal distances. In *International Conference on Social Computing (SocialCom)*, 2011.
64. Freédéric Cupillard, Francois Brémond, Monique Thonnat, Inria Sophia Antipolis, and Orion Group. Tracking groups of people for video surveillance. 2001.
65. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society.
66. D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):224 –227, april 1979.
67. M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas. Learning where to attend with deep architectures for image tracking. *Neural Computation*, 2012. in print.
68. P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 304 –311, june 2009.
69. A Doucet, N de Freitas, and N Gordon. Introduction to sequential Monte Carlo methods. In A Doucet, N de Freitas, and N J Gordon, editors, *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.
70. Arnaud Doucet, Nando De Freitas, and Neil Gordon. *Sequential Monte Carlo methods in practice*. 2001.
71. G. J. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In *International Conference on Face & Gesture Recognition*, 1998.
72. P. Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384, 1993.
73. A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151 – 1163, 2002.

74. Tom Erez, Julian J. Tramper, William D. Smart, and Stan C. A. M. Gielen. A pomdp model of eye-hand coordination. In Wolfram Burgard and Dan Roth, editors, *25th Conference on Artificial Intelligence (AAAI 2011)*, pages 952–957. AAAI Press, 2011.
75. A. Ess, K. Schindler, B. Leibe, and L. Van Gool. Improved multi-person tracking with active occlusion handling. In *Workshop on People Detection and Tracking*, 2009.
76. M. Farenzena, L. Bazzani, V. Murino, and M. Cristani. Towards a subject-centered analysis for automated video surveillance. In *Proceedings of the 15th International Conference on Image Analysis and Processing (ICIAP)*, pages 481–489. Springer-Verlag, 2009.
77. M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2360 –2367, June 2010.
78. M. Farenzena, A. Tavano, L. Bazzani, D. Tosato, G. Pagetti, G. Menegaz, V. Murino, and M. Cristani. Social interaction by visual focus of attention in a three-dimensional environment. In *Workshop on Pattern Recognition and Artificial Intelligence for Human Behavior Analysis at AI*IA*, 2009.
79. M. Feldmann, D. Fränken, and W. Koch. Tracking of extended objects and group targets using random matrices. *IEEE Transactions on Signal Processing*, 59(4):1409–1420, 2011.
80. P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Cascade Object Detection with Deformable Part Models. CVPR, 2010.
81. P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.
82. Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2010.
83. Paul Fieguth and Demetri Terzopoulos. Color-based tracking of heads and other mobile objects at video frame rates. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1997.
84. M. Figueiredo and A.K. Jain. Unsupervised learning of finite mixture models. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.
85. Per-Erik Forssén. Maximally stable colour regions for recognition and matching. 2007.
86. L. Freeman. Social networks and the structure experiment. In *Research Methods in Social Network Analysis*, pages 11–40, 1989.
87. Y Freund and D Haussler. Unsupervised learning of distributions of binary vectors using 2-layer networks. In *Advances in Neural Information Processing Systems 4 (NIPS 4)*, pages 912–919. Morgan Kaufman Publishers, 1991.
88. Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.
89. J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *The annals of statistics*, 28(2):337–374, 2000.
90. R. Gaborski, V. Vainganker, V. Chaoji, A. Teredesai, and A. Tentler. Detection of inconsistent regions in video streams. In *Proc. SPIE Human Vision and Electronic Imaging*. Citeseer, 2004.
91. D. Gao, V. Mahadevan, and N. Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. *Advances in neural information processing systems*, 20, 2007.

92. 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, 2009.
93. W. Ge, R. T. Collins, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints), 2011.
94. G Gennari and G.D. Hager. Probabilistic data association methods in visual tracking of groups. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.
95. N. Gheissari, T. B. Sebastian, P. H. Tu, J. Rittscher, and R. Hartley. Person reidentification using spatiotemporal appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1528–1535, 2006.
96. R. Gherardi, M. Farenzena, and A. Fusiello. Improving the efficiency of hierarchical structure-and-motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1594 –1600, june 2010.
97. B. Girard and A. Berthoz. From brainstem to cortex: Computational models of saccade generation circuitry. *Progress in Neurobiology*, 77(4):215 – 251, 2005.
98. A. Gning, L. Mihaylova, S. Maskell, Sze Kim Pang, and S. Godsill. Group object structure and state estimation with evolving networks and monte carlo methods. *Signal Processing, IEEE Transactions on*, 59(4):1383 –1396, april 2011.
99. Jacqueline P. Gottlieb, Makoto Kusunoki, and Michael E. Goldberg. The representation of visual salience in monkey parietal cortex. *Nature*, 391:481–484, 1998.
100. Helmut Grabner, Christian Leistner, and Horst Bischof. Semi-supervised on-line boosting for robust tracking. In *European Conference on Computer Vision*, pages 234–247, Berlin, Heidelberg, 2008. Springer-Verlag.
101. D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007.
102. Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, Jul 2008.
103. Edward T Hall. *The Hidden Dimension.* 1966.
104. 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 *IEEE International Conference on Distribuited Smart Cameras*, pages 1–6, 2008.
105. G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.
106. Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.
107. Geoffrey E. Hinton. A practical guide to training restricted boltzmann machines. TR 2010-003, University of Toronto, 2010.
108. S. Hongeng and R. Nevatia. Large-scale event detection using semi-hidden markov models. In *IEEE International Conference on Computer Vision*, volume 2, 2003.
109. M. Hu, S. Ali, and M. Shah. Learning motion patterns in crowded scenes using motion flow field. In *International Conference on Pattern Recognition*, pages 1–5, 2008.
110. Chang Huang, Bo Wu, and Ramakant Nevatia. Robust object tracking by hierarchical association of detection responses. In *European Conference on Computer Vision*, pages 788–801, Berlin, Heidelberg, 2008. Springer-Verlag.
111. M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *ECCV*, pages 343–356, 1996.

112. M. Isard and A. Blake. Condensation: Conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29:5–28, 1998.
113. M. Isard and J. MacCormick. Bramble: A bayesian multiple-blob tracker. 2001.
114. L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254 –1259, 1998.
115. B. Jabarin, J. Wu, R. Vertegaal, and L. Grigorov. Establishing remote conversations through eye contact with physical awareness proxies. In *CHI '03 extended abstracts*, 2003.
116. Omar Javed, Khurram Shafique, Zeeshan Rasheed, and Mubarak Shah. Modeling inter-camera space-time and appearance relationships for tracking accross non-overlapping views. *Computer Vision and Image Understanding*, 109:146–162, 2007.
117. Omar Javed and Mubarak Shah. *Automated Multi-Camera Surveillance: Algorithms and Practice*. 2008.
118. N. Jojic, B. Frey, and A. Kannan. Epitomic analysis of appearance and shape. *International Conference on Computer Vision*, 1:34–41, Oct 2003.
119. N. Jojic, A. Perina, M. Cristani, V. Murino, and B. Frey. Stel component analysis: Modeling spatial correlations in image class structure. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2044–2051, 2009.
120. S. Julier and J. Uhlmann. A new extension of the kalman filter to nonlinear systems. In *Int. Symp. Aerospace/Defense Sensing, Simul. and Controls, Orlando, FL*, 1997.
121. T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE Transactions on Communications*, 15(1):52–60, 1967.
122. Z. Kalal, J. Matas, and K. Mikolajczyk. P-n learning: Bootstrapping binary classifiers by structural constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 49 –56, june 2010.
123. Z. Kalal, K. Mikolajczyk, and J. Matas. Face-tld: Tracking-learning-detection applied to faces. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 3789–3792. IEEE, 2010.
124. R. E. Kalman. A new approach to linear filtering and prediction problems. *Tran. of the ASME Journal of Basic Engineering*, (82 (Series D)):35–45, 1960.
125. Christopher Kanan and Garrison W. Cottrell. Robust classification of objects, faces, and flowers using natural image statistics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, pages 2472–2479, 2010.
126. R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 319–336, 2009.
127. K. Kavukcuoglu, M.A. Ranzato, R. Fergus, and Yann Le-Cun. Learning invariant features through topographic filter maps. In *CVPR*, pages 1605–1612, 2009.
128. Z. Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(11):1805–1819, 2005.
129. 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 transactions on pattern analysis and machine intelligence*, 28(12):1960–1972, 2006.
130. M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. *IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

131. M. L. Knapp and J. A. Hall. *Nonverbal communication in human interaction.* Wadsworth Pub Co, 2009.
132. C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 4(4):219–27, 1985.
133. W. Kohler. *The task of Gestalt psychology.* Princeton NJ, 1969.
134. Urs Köster and Aapo Hyvärinen. A two-layer ICA-like model estimated by score matching. In *ICANN*, pages 798–807, 2007.
135. Cheng-Hao Kuo, Chang Huang, and R. Nevatia. Multi-target tracking by on-line learned discriminative appearance models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 685 –692, 2010.
136. J. S. Kwon and K. M. Lee. Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1208–1215, 2009.
137. A. Lablack and C. Djeraba. Analysis of human behaviour in front of a target scene. In *IEEE International Conference on Pattern Recognition*, pages 1–4, 2008.
138. C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.
139. Tian Lan, Yang Wang, Weilong Yang, and Greg Mori. Beyond actions: Discriminative models for contextual group activities. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.
140. S. H. R Langton, R. J. Watt, and V. Bruce. Do the eyes have it? cues to the direction of social attention. *Trends in Cognitive Neuroscience*, 4(2):50–58, 2000.
141. O. Lanz, R. Brunelli, P. Chippendale, M. Voit, and R. Stiefelhagen. *Extracting Interaction Cues: Focus of Attention, Body Pose, and Gestures*, pages 87–93. Springer, 2009.
142. Oswald Lanz. Approximate bayesian multibody tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(9):1436–1449, 2006.
143. Hugo Larochelle and Geoffrey E. Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine. In *Neural Information Processing Systems*, 2010.
144. B. Lau, K.O. Arras, and W. Burgard. Multi-model hypothesis group tracking and group size estimation. *International Journal of Social Robotics*, 2(1):19–30, 2010.
145. Y. LeCun and C. Cortes. The MNIST database of handwritten digits. *NEC Research Institute, http://yann. lecun. com/exdb/mnist/index. html*.
146. H. Lee, R. Grosse, R. Ranganath, and A.Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, 2009.
147. A. Levinshtein, S. Dickinson, and C. Sminchisescu. Multiscale Symmetric Part Detection and Grouping. In *International Conference on Computer Vision*, 2009.
148. Peihua Li, Tianwen Zhang, and Bo Ma. Unscented kalman filter for visual curve tracking. *Image and Vision Computing*, 22(2):157 – 164, 2004. Statistical Methods in Video Processing.
149. Stan Z. Li, Long Zhu, ZhenQiu Zhang, Andrew Blake, HongJiang Zhang, and Harry Shum. Statistical learning of multi-view face detection. In *Proceedings of the 7th European Conference on Computer Vision*, ECCV '02, pages 67–81, London, UK, UK, 2002. Springer-Verlag.
150. X Li, W Hu, Z Zhang, X Zhang, M Zhu, and J Cheng. Visual tracking via incremental log-euclidean riemannian subspace learning. *The International Journal of Robotics Research*, pages 1–8, 2008.
151. X Rong Li and Y Bar-Shalom. Tracking in clutter with nearest neighbor filters: analysis and performance. *IEEE Trans. on Aerospace and Electronic Systems*, 32(3):995–1010, 1996.

152. Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2953–2960, 2009.
153. Jongwoo Lim, David Ross, Ruei sung Lin, and Ming hsuan Yang. Incremental learning for visual tracking. In *Conference on Advances in Neural Information Processing Systems*, pages 793–800. MIT Press, 2004.
154. Weiyao Lin, Ming-Ting Sun, R. Poovendran, and Zhengyou Zhang. Group event detection with a varying number of group members for video surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(8):1057 –1067, aug. 2010.
155. Wen-Chieh Lin and Yanxi Liu. A lattice-based mrf model for dynamic near-regular texture tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(5):777–792, 2007.
156. Zhe Lin and Larry S. Davis. Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In *International Symposium on Advances in Visual Computing*, pages 23–34, Berlin, Heidelberg, 2008. Springer-Verlag.
157. GuoJun Liu, XiangLong Tang, JianHua Huang, JiaFeng Liu, and Da Sun. Hierarchical model-based human motion tracking via unscented kalman filter. *Computer Vision, IEEE International Conference on*, 0:1–8, 2007.
158. Rong Liu, Jian Cheng, and Hanqing Lu. A robust boosting tracker with minimum error bound in a co-training framework. In *International Conference on Computer Vision*, pages 1459–1466, Sep 2009.
159. X. Liu, N. Krahnstoever, Y. Ting, and P. Tu. What are customers looking at? In *Advanced Video and Signal Based Surveillance*, pages 405–410, 2007.
160. E. Maggio, E. Piccardo, C. Regazzoni, and A. Cavallaro. Particle phd filtering for multi-target visual tracking. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 1101–1104, 2007.
161. E. Maggio, F. Smerladi, and A. Cavallaro. Combining colour and orientation for adaptive particle filter-based tracking. In *British Machine Vision Conference*, pages xx–yy, 2005.
162. D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–205–II–210 Vol.2, 2004.
163. Jorge S Marques, Pedro M Jorge, Arnaldo J Abrantes, and J M Lemos. Tracking groups of pedestrians in video sequences. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop*, volume 9, pages 101–101, Jun 2003.
164. David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Henry Holt and Co., Inc., New York, NY, USA, 1982.
165. Y. Matsumoto, T. Ogasawara, and A. Zelinsky. Behavior recognition based on head-pose and gaze direction measurement. In *Proc. Int'l Conf. Intelligent Robots and Systems*, volume 4, pages 2127–2132, 2002.
166. I. Matthews, T. Ishikawa, and S. Baker. The template update problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 810–815, 2004.
167. Iain Matthews and Simon Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(1):135 – 164, November 2004.
168. T. Mauthner, M/ Donoser, and H. Bischof. Robust tracking of spatial related components. In *International Conference on Pattern Recognition*, pages 1–4, Dec 2008.
169. Stephen J. Mckenna, Sumer Jabri, Zoran Duric, Harry Wechsler, and Azriel Rosenfeld. Tracking groups of people. *Computer Vision and Image Understanding*, pages 42–56, 2000.

170. Bruce L. McNaughton, Francesco P. Battaglia, Ole Jensen, Edvard I. Moser, and May-Britt Moser. Path integration and the neural basis of the 'cognitive map'. *Nature Reviews Neuroscience*, 7(8):663–678, 2006.

171. B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19:696–710, July 1997.

172. Kevin Patrick Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley, 2002.

173. Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31:607–626, April 2009.

174. Jiri Najemnik and Wilson S. Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, March 2005.

175. C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. *Pattern Recognition*, 36(9), 2003.

176. J. D. Nelson, C. R. M. McKenzie, G. W. Cottrell, and T. J. Sejnowski. Experience matters: Information acquisition optimizes probability gain. *Psychological science*, 21(7):960–969, 2010.

177. Bingbing Ni, Shuicheng Yan, and Ashraf A. Kassim. Recognizing human group activities with localized causalities. In *CVPR'09*, pages 1470–1477, 2009.

178. K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *European Conference on Computer Vision*, pages Vol I: 28–39, 2004.

179. B. A. Olshausen, C. H. Anderson, and D. C. Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, 13(11):4700, 1993.

180. B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

181. Bruno A. Olshausen, Charles H. Anderson, and David C. Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13:4700–4719, 1993.

182. S. Paisitkriangkrai, C. H. Shen, and J. Zhang. Performance evaluation of local features in human classification and detection. *Computer Vision, Institution of Engineering and Technology*, 2(4):236–246, 2008.

183. H. Palaio and J.P. Batista. A region covariance embedded in a particle filter for multi-objects tracking. In *VS08*, 2008.

184. P. Pan and D. Schonfeld. Dynamic proposal variance and optimal particle allocation in particle filtering for video tracking. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(9):1268–1279, 2008.

185. J. Panero and M. Zelnik. *Human Dimension and Interior Space : A Source Book of Design*. 1979.

186. Sze Kim Pang, Jack Li, and Simon Godsill. Models and algorithms for detection and tracking of coordinated groups. In *Symposium of image and Signal Processing and Analisys*, 2007.

187. Nikos Paragios and Rachid Deriche. Geodesic active regions and level set methods for supervised texture segmentation. *International Journal of Computer Vision*, 46:223–247, February 2002.

188. Sangho Park and Mohan M. Trivedi. Multi-person interaction and activity analysis: a synergistic track- and body-level analysis framework. *Mach. Vision Appl.*, 18:151–166, May 2007.

189. S. Pellegrini, A. Ess, M. Tanaskovic, and L. Van Gool. Wrong turn - no dead end: A stochastic pedestrian motion model. In *Computer Vision and Pattern Recognition*

*Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 15 –22, june 2010.

190. Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *International Conference on Computer Vision*, 2009.
191. Stefano Pellegrini, Andreas Ess, and Luc Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *Proceedings of the 11th European conference on Computer vision: Part I*, ECCV'10, pages 452–465, Berlin, Heidelberg, 2010. Springer-Verlag.
192. A. Pentland and S. Pentland. *Honest signals: how they shape our world*. The MIT Press, 2008.
193. Alex Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:107–119, January 2000.
194. Patrick Pérez, Carine Hue, Jaco Vermaak, and Michel Gangnet. Color-based probabilistic tracking. In *European Conference on Computer Vision*, pages 661–675, London, UK, 2002. Springer-Verlag.
195. N. T. Pham, W. M. Huang, and S. H. Ong. Probability hypothesis density approach for multi-camera multi-object tracking. In *Asian Conference on Computer Vision*, pages I: 875–884, 2007.
196. Fatih Porikli, Oncel Tuzel, and Peter Meer. Covariance tracking using model update based on lie algebra. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 728–735, Washington, DC, USA, 2006. IEEE Computer Society.
197. Eric O. Postma, H. Jaap van den Herik, and Patrick T. W. Hudson. SCAN: A scalable model of attentional selection. *Neural Networks*, 10(6):993 – 1015, 1997.
198. F. P. Preparata and M. I. Shamos. *Computational Geometry. An Introduction.* 1985.
199. B. Prosser, W. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *British Machine Vision Conference*, 2010.
200. G. Psathas. *Conversation analysis: The study of talk-in-interaction.* Sage Publications, Inc, 1995.
201. A. Rahimi, B. Dunagan, and T. Darrel. Simultaneous calibration and tracking with a network of non-overlapping sensors. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 187–194, 2004.
202. Marc'Aurelio Ranzato, Alex Krizhevsky, and Geoffrey E. Hinton. Factored 3-way restricted boltzmann machines for modeling natural images. In *International Conference on Artificial Intelligence and Statistics*, 2010.
203. Marc'Aurelio Ranzato, Joshua Susskind, Volodymyr Mnih, and Geoffrey Hinton. On deep generative models with applications to recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.
204. C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning.* Adaptive computation and machine learning. MIT Press, 2006.
205. Christopher Rasmussen and Gregory D Hager. Probabilistic data association methods for tracking multiple and compound visual objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23:560–576, 2000.
206. D. Reisfeld, H. J. Wolfson, and Y. Yeshurun. Context-free attentional operators: The generalized symmetry transform. *International Journal of Computer Vision*, 14(2):119–130, 1995.
207. R. A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7(1):17–42, 2000.
208. V. P. Richmond, J. C. McCroskey, and S. K. Payne. *Nonverbal behavior in interpersonal relations.* Allyn and Bacon, 2000.

209. T. Riklin-Raviv, N. Sochen, and N. Kiryati. On symmetry, perspectivity, and level-set-based segmentation. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 31(8):1458–1471, 2009.

210. Neil Robertson and Ian Reid. *Estimating Gaze Direction from Low-Resolution Faces in Video.* 2006.

211. Mikel Rodriguez, Saad Ali, and Takeo Kanade. Tracking in unstructured crowded scenes. In *International Conference on Computer Vision*, 2009.

212. M. G. P. Rosa. Visual maps in the adult primate cerebral cortex: Some implications for brain development and evolution. *Brazilian Journal of Medical and Biological Research*, 35:1485 – 1498, 2002.

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

214. Daniel Rowe, Ignasi Rius, Jordi Gonzàlez, and Juan José Villanueva. Improving tracking by handling occlusions. In *Advances in Pattern Recognition*, pages 384–393, 2005.

215. R. J. Rummel. *Understanding conflict and war.* Sage Publications, 1981.

216. A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof. On-line random forests. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1393–1400. IEEE, 2009.

217. J. Santner, C. Leistner, A. Saffari, T. Pock, and H. Bischof. PROST: Parallel robust online simple tracking. 2010.

218. Lawrence K Saul and Michael I Jordan. Mixed memory markov models: Decomposing complex stochastic processes as mixtures of simpler ones. *Machine Learning*, 37(1):75–87, 1999.

219. R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Mach. Learn.*, 37:297–336, 1999.

220. A. E. Scheflen. The significance of posture in communication systems. *Communication Theory*, page 293, 2007.

221. K. R. Scherer. *Personality markers in speech.* Cambridge Univ. Press, 1979.

222. W. R. Schwartz and L. S. Davis. Learning discriminative appearance-based models using partial least squares. In *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing*, 2009.

223. P. Scovanner and M. F. Tappen. Learning pedestrian dynamics from the real world. In *International Conference on Computer Vision*, pages 381–388, 2009.

224. P. Sebastian, Yap Vooi Voon, and R. Comley. The effect of colour space on tracking robustness. pages 2512–2516, 2008.

225. David Serby, Esther Koller-Meier, and Luc Van Gool. Probabilistic object tracking using multiple features. pages 184–187, 2004.

226. J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *Proceedings of the British Machine Vision Conference*, 2006.

227. K. Smith, S. O. Ba, J. M. Odobez, and D. Gatica-Perez. Tracking the visual focus of attention for a varying number of wandering people. *IEEE transactions on pattern analysis and machine intelligence*, 30(7):1212–1229, 2008.

228. Kevin Smith, Daniel Gatica-Perez, Jean-Marc Odobez, and Sileye Ba. Evaluating multi-object tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 36, 2005.

229. P. Smith, M. Shah, and N. da Vitoria Lobo. Determining driver visual attention with one camera. *IEEE Transactions on Intelligent Transportation Systems*, 4(4):205–218, 2003.

230. Paul Smolensky. Information Processing in Dynamical Systems: Foundations of Harmony Theory. volume 1, chapter 6, pages 194–281. MIT Press, Cambridge, 1986.

231. N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM Transactions on Graphics*, volume 25, pages 835–846. ACM, 2006.
232. C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. volume 2, pages 252–259, 1999.
233. R. Stiefelhagen, R. Bowers, and J. Fiscus, editors. *Multimodal Technologies for Perception of Humans: International Evaluation Workshops on Classification of Events, Activities and Relationships 2007*. Springer-Verlag, Berlin, Heidelberg, 2008.
234. R. Stiefelhagen and J. Garofolo, editors. *Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities and Relationships 2006*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.
235. R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13:928–938, 2002.
236. Rainer Stiefelhagen, Michael Finke, Jie Yang, and Alex Waibel. From gaze to focus of attention. In *Visual Information and Information Systems*, pages 761–768, 1999.
237. Deqing Sun, Erik Sudderth, and Michael Black. Layered image motion with explicit occlusions, temporal consistency, and depth ordering. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2226–2234. 2010.
238. K. Swersky, D. Buchman, B. M. Marlin, and N. de Freitas. On autoencoders and score matching for energy based models. *International Conference in Machine Learning*, 2011.
239. Graham W. Taylor, Leonid Sigal, David J. Fleet, and Geoffrey E. Hinton. Dynamical binary latent variable models for 3d human pose tracking. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:631–638, 2010.
240. A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological review*, 113(4):766, 2006.
241. D. Tosato, M. Farenzena, M. Spera, M. Cristani, and V. Murino. Multi-class classification on riemannian manifolds for video surveillance. In *IEEE European Conference on Computer Vision*, 2010.
242. O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. *Proceedings of the European Conference on Computer Vision*, pages 589–600, 2006.
243. Oncel Tuzel, Fatih Porikli, and Peter Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30:1713–1727, October 2008.
244. Raquel Urtasun, David J. Fleet, and Pascal Fua. 3d people tracking with gaussian process dynamical models. In *EEE Conference on Computer Vision and Pattern Recognition*, pages 238–245, 2006.
245. Rudolph van der Merwe, Nando de Freitas, Arnaud Doucet, and Eric Wan. The unscented particle filter. In *Advances in Neural Information Processing Systems 13*, November 2001.
246. Namrata Vaswani, Amit Roy Chowdhury, and Rama Chellappa. Activity recognition using the dynamics of the configuration of interacting objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–640, 2003.
247. Cor J. Veenman, C. J. Veenman, Marcel J. T. Reinders, and Eric Backer. Resolving motion correspondence for densely moving points. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:54–72, 2001.
248. Jaco Vermaak, Simon J. Godsill, and Patrick Pérez. Monte carlo filtering for multi-target tracking and data association. *IEEE Transactions on Aerospace and Electronic Systems*, 41:309–332, 2004.

249. P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103, 2008.
250. A. Vinciarelli, M. Pantic, and H. Bourlard. Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759, 2009.
251. A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social signals, their function, and automatic analysis: a survey. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 61–68, New York, NY, USA, 2008. ACM.
252. M. Viola, M.J. Jones, and P. Viola. Fast multi-view face detection. In *Proc. of Computer Vision and Pattern Recognition*. Citeseer, 2003.
253. Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision*, 2002.
254. M. Voit and R. Stiefelhagen. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In *Proceedings of the 10th international conference on Multimodal interfaces*, ICMI '08, pages 173–180, New York, NY, USA, 2008. ACM.
255. A. Waibel, T. Schultz, M. Bett, M. Denecke, R. Malkin, I. Rogina, and R. Stiefelhagen. SMaRT: the Smart Meeting Room task at ISL. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 752–755, 2003.
256. M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.
257. Gang Wang, Andrew Gallagher, Jiebo Luo, and David Forsyth. Seeing people in social context: Recognizing people and social relationships. In *European Conference on Computer Vision*, pages 169–182, 2010.
258. J. Q. Wang and Y. S. Yagi. Switching local and covariance matching for efficient object tracking. In *International Conference on Pattern Recognition*, pages 1–4, 2008.
259. X. Wang, G. Doretto, T. B. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In *International Conference on Computer Vision*, pages 1–8, 2007.
260. Xiaogang Wang, Xiaoxu Ma, and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31:539–555, March 2009.
261. Ya-Dong Wang, Jian-Kang Wu, Ashraf A Kassim, and Wei-Min Huang. Tracking a variable number of human groups in video using probability hypothesis density. 2006.
262. R. M. Warner and D. B. Sugarman. Attributions of personality based on physical appearance, speech, and handwriting. *Journal of Personality and Social Psychology*, 50(4):792, 1986.
263. M. Welling, M. Rosen-Zvi, and G. Hinton. Exponential family harmoniums with an application to information retrieval. *NIPS*, 17:1481–1488, 2005.
264. S. Whittaker, D. Frohlich, and O. Daly-Jones. Informal workplace communication: what is it like and how might we support it? In *CHI '94*, page 208, 1994.
265. 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.
266. B. Wu and R. Nevatia. Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection. In *Proceedings of the International Conference of Computer Vision and Pattern Recognition*, 2008.

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

268. Bo Wu, Haizhou Ai, Chang Huang, and Shihong Lao. Fast rotation invariant multi-view face detection based on real adaboost. In *Proceedings of the Sixth IEEE international conference on Automatic face and gesture recognition*, FGR' 04, pages 79–84, Washington, DC, USA, 2004. IEEE Computer Society.

269. J. L. Xing, H. Z. Ai, and S. H. Lao. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1200–1207, 2009.

270. K. Yamaguchi, A.C. Berg, L.E. Ortiz, and T.L. Berg. Who are you with and where are you going? In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.

271. Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13, 2006.

272. Alper Yilmaz, Xin Li, and Mubarak Shah. Contour-based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 26:1531–1536, November 2004.

273. Z. Yin and R. T. Collins. Object tracking and detection after occlusion via numerical hybrid local and global mode-seeking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

274. Qian Yu and Gerard Medioni. Multiple-target tracking by spatiotemporal monte carlo markov chain data association. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(12):2196–2210, 2009.

275. W. S. Zeng, Shaogang Gong, and T. Xiang. Associating groups of people. In *British Conference on Machine Vision*, 2009.

276. L. Zhang, M. H. Tong, and G. W. Cottrell. Sunday: Saliency using natural statistics for dynamic analysis of scenes. In *Proceedings of the 31st Annual Cognitive Science Conference, Amsterdam, Netherlands*. Citeseer, 2009.

277. Song Chun Zhu and Alan Yuille. Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:884–900, 1996.

278. I. Zuriarrain, F. Lerasle, N. Arana, and M. Devy. An mcmc-based particle filter for multiple person tracking. In *International Conference on Pattern Recognition*, 2008.