Analyzing Groups: a Social Signaling Perspective

Loris Bazzani, Marco Cristani, Giulia Paggetti, Diego Tosato, Gloria Menegaz, Vittorio Murino

Abstract This chapter introduces some basic methods to deal with groups of people in surveillance settings. Recently, modeling groups has become a very active trend for video surveillance researchers. Our solution is proper of the recently forged field of social signaling, since it embeds notions of social psychology into computer vision techniques, offering a novel research perspective for the video surveillance community. In particular, we present methods to discover and track groups of people, and to infer what is the focus of attention of each person, that is, we estimate the portion of a scene that is frequently observed by people. Each method we present is evaluated in an experimental section on real scenario, that gives a clear idea of its performance and potentialities.

1 Introduction

Recently, researchers in surveillance shifted the attention from the monitoring of a single person in a camera-monitored environment to that of groups: this novel level of abstraction provides event descriptions which are semantically more meaningful, highlighting barely visible relational connections among people. 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 among humans. Actually, to the best of our knowledge, all

Loris Bazzani, Giulia Paggetti, Diego Tosato, Gloria Menegaz
University of Verona, Verona, Italy, e-mail: name.surname@univr.it

Marco Cristani
University of Verona, Verona, Italy, and
Istituto Italiano di Tecnologia (IIT), Genova, Italy, e-mail: marco.cristani@iit.it

Vittorio Murino
Istituto Italiano di Tecnologia (IIT), Genova, Italy, e-mail: vittorio.murino@iit.it

the work that deal with groups assumes that people are simple points on a plane [23, 32, 78, 53, 48, 84, 39] that in some cases may obey to physical laws of attraction and repulsion [54, 66]. None of them considers that working on groups implies to focus on the analysis of the human behavior – a process subject 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 signaling and affective computing [80], that typically neglect scenarios relevant to surveillance and monitoring.

Social signaling and computer vision are tightly intertwined. In our context, they attempts to discover social interactions using statistical analysis of spatial-orientational arrangements that are relevant in a social psychology sense. 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[1]: physical appearance (attractiveness, clothes, ornaments, somatotype, etc.) [59, 86], vocal behavior (everything else than words in speech) [58, 65], face and eyes behavior (expressions, gaze, head pose, etc.) [10, 13], gestures and postures (hand and body movements, conscious and unconscious gestures, orientation with respect to others, etc.) [55, 64], space and environment (mutual distances, spatial organization of people, territoriality, geometric constraints) [30, 59]. In this context, social interactions are here intended as the acts, actions, or practices of two or more peoepl mutually oriented versus each other, that is, every behaviors affecting or considering others' subjective experiences or intentions [61]. 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 among individuals (or groups) that modify their actions and reactions by their interaction partner(s) are social interactions.

The methods presented here 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 of people 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 in a typical surveillance scenario, where the input of the method is just a video. For this reason, we set up a definition of group that considers some social signaling cues coming from the code of face and eyes behavior and the space and environment component. In this work, we consider several aspect of groups: 1) 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), 2) how a moving group can be tracked (its *evolution*), and 3) which basic activities are carried out by their components in terms of interactions between the humans and the environment. In particular, we detect the regions of the environment where the attention of humans is more focused.

[1] For a complete review of social signaling, please read [79].

The *birth of a group* or its initialization (and, consequently, its break-up) can be performed in two ways. The first assumes that individuals 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 initialization 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.

The *evolution of a group* considers a group while it is moving. In this case, we present a tracking approach which embeds the knowledge of the states of the single individuals and the state of a group for providing a robust group localization.

Moreover, we present the idea of how people interact with an environment through an *interest map*, *i.e.*, a map that highlights the part of a scene more considered 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 scholars 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 feature we propose 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 2 details the basic elements proposed by the computer vision community, and used here both as low-level information and as a basis to understand the techniques illustrated in the subsequent Sections. The core of the chapter is represented by Sections 3, 4 and 5, describing approaches to initialize the groups, to track groups and to create interest maps of a monitored setting, respectively. Finally, conclusions are drawn in Section 6.

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 in a bottom-up way for extracting low-level features from videos, useful to allow high-level inference. First, all the people in the camera-monitored environment are localized, by exploting a tracker that provides the trajectory of each person. When the position is estimated, a head orientation method computes the pose of the head of each person, and the subjective field of view (i.e., the view frustum) is initializated by exploiting the calibration of the camera. Thus, the basic components used by the proposed architecture are: a multi-target tracker (see Sec. 2.1), a head pose computation method (see Sec. 2.2), and the subjective view frustum estimation (see Sec. 2.3).

2.1 Tracking with Particle Filters

Many multi-target tracking techniques have been successfully proposed in literature from the Kalman filter and its extensions [28, 27] to the more recent Probability Hypothesis Density filter [41] and particle filter (PF) [12, 51]. The success of a tracking algorithm depends on several factors: the strategy used for tracking, the data association approach, the appearance model, occlusion modeling, and so on. Among the realm of the tracking strategies, an important role is played by the particle filtering [12]: the general framework and the simplicity of the method make it one of the most used methods for tracking in the last years.

Particle filtering offers a probabilistic framework for recursive dynamic state estimation. The approach was born originally for single-target tracking [24], then later it was extended to a multi-target tracking scenario [25]. 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, which is (in general) exponential in the number of targets to track. Recently, an interesting yet general solution has been proposed in [35]. Here, the Hybrid Joint-Separable (HJS) filter is introduced, that maintains a linear relationship between the number of targets and the number of particles. In addition, an occlusion model has been proposed exploiting the camera calibration.

From a Bayesian perspective, the single object tracking problem aims at recursively calculating the posterior distribution $p(x_t|z_{1:t})$ by exploiting the Chapman-Kolmogorov equation, where x_t is the current state of the target (*e.g.*, its position and its scale), z_t is the current measurement (*e.g.*, the current frame), and $x_{1:t}$ and $z_{1:t}$ are the states and the measurements up to time t, respectively. In formulae:

$$p(x_t|z_{1:t}) \propto p(z_t|x_t) \int p(x_t|x_{t-1}) p(x_{t-1}|z_{1:t-1}) dx_{t-1} \quad (1)$$

The equation is fully specified by an initial distribution $p(x_0|z_0) = p(x_0)$, the dynamical model $p(x_t|x_{t-1})$, and the observation model $p(z_t|x_t)$. Particle filtering (PF) approximates the posterior distribution by a set of N weighted particles, *i.e.*, $\{(x_t^{(n)}, w_t^{(n)})\}_{n=1}^N$; a large weight $w_t^{(n)}$ mirrors a state $x_t^{(n)}$ with high posterior probability. In this way, the integral in Eq. 1 has not to be analytically solved, and, instead, the posterior at time $t-1$ is sampled, defining a set of state hypotheses (the particles) that evolve according to the dynamical model $p(x_t|x_{t-1})$ (the prediction step), and which is evaluated via $p(z_t|x_t)$ (the observation step). The parameter N is manually set here, but there exist techniques for estimating the optimal number of particles that minimizes some measure of tracking distortion [51].

HJS filter [35] is an extension of the PF for multiple targets. Defining $\mathbf{x}_t = \{x_t^1, x_t^2, \ldots, x_t^K\}$ the joint state (the ensemble of the K individual states), HJS adopts the approximation $p(\mathbf{x}_t|z_{1:t}) \approx \prod_k p(x_t^k|z_{1:t})$, that is, the joint posterior could be approximated via the product of its marginal components (k indexes the individual targets). The dynamics and the observation models of HJS are expressed as follows:

$$p(x_t^k|x_{t-1}^k) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}^{\neg k}|z_{1:t-1}) d\mathbf{x}_{t-1:t}^{\neg k} \quad (2)$$

$$p(z_t|x_t^k) = \int p(z_t|\mathbf{x}_t) p(\mathbf{x}_t^{\neg k}|z_{1:t-1}) d\mathbf{x}_t^{\neg k} \quad (3)$$

where the apex $^{\neg k}$ addresses all the targets but the kth. These equations encode an intuitive strategy, i.e., that both the dynamics and the observation phases of the kth target lie upon the consideration of a joint dynamical model $p(\mathbf{x}_t|\mathbf{x}_{t-1}) \approx p(\mathbf{x}_t) \prod_k q(x_t^k|x_{t-1}^k)$ and observation model $p(z_t|\mathbf{x}_t)$. The joint dynamical model, through the prior $p(\mathbf{x}_t)$, avoids that multiple targets with single motion described by $q(x_t^k|x_{t-1}^k)$ collapse in a single location, and the joint observation model considers that the visual appearance of a single target might be occluded by another object, acting as a z-buffer. The two models are weighted by posterior distributions that essentially promote trusted joint objects configurations (not considering the kth object).

The observational model $p(z_t|\mathbf{x}_t)$ quantifies the likelihood of the single measure z_t given the state $\mathbf{x}_t$, considering inter-objects occlusions. It is built upon the representation of the targets, that here are constrained to be human beings. The human body is represented by its three components: head, torso and legs. The observational model works by evaluating a separate appearance score for each object (summing then the contribute of the single parts). This score is encoded by a distance between the histograms of the model and the hypothesis (a sample), and it 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 $\{x_p\}_{p=1}^{N \cdot K}$ in an incremental visit procedure on the ground floor. The hypothesis nearest to the camera is evaluated first. Its presence determines

an occluding cone in the scene, with an associated confidence that depends on the observational likelihood achieved. Parts of other objects farther from the camera that fall in the occlusion cone are considered less important in their observational likelihood computation. The process of map building is iterated going deeper and deeper in the scene.

In formulae, the observation model is defined as

$$p(z_t|x_p) \propto \exp\left(-\frac{\mathrm{fc}_p + \mathrm{bc}_p}{2\,\sigma^2}\right), \tag{4}$$

where fc_p is the foreground term, *i.e.*, the likelihood that an object matches to the model considering the un-occluded parts, and bc_p, the background term, accounts for the occluded parts of an object. For more details, readers may refer to [35].

2.2 Head Orientation Estimation

Head orientation estimation is becoming an important computer vision application. There are several diverse approaches present in the literature: a recent review can be found in [47], where a performance analysis of different methods is presented, and a list of the commonly used dataset for head pose estimation is shown. The CLEAR workshops are important events for the head pose estimation community, and several important approaches can be found in the related proceedings [72, 71]. 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 [37, 81, 90, 77, 88, 50]. Boosting [18, 63, 19] is a remarkable, highly customizable way to create 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 [89] for an updated list), covariance features [76] have been exploited as powerful descriptors of pedestrians [77, 88], and their effectiveness has been explicitly investigated in a comparative study [50]. When injected in boosting systems [77, 88, 50, 75], 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. In this chapter, we present in details the method proposed in [75].

The tracker provides the location of the head and the feet for each person in each frame. As for the head approximate position, 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.

For the multi-class classification, we boost regression trees [6, 75], 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 overtraining of the regression tree, we establish as stopping rule a minimal number τ of observations per tree leaf, that is experimentally estimated (see Sec. 3.2).

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

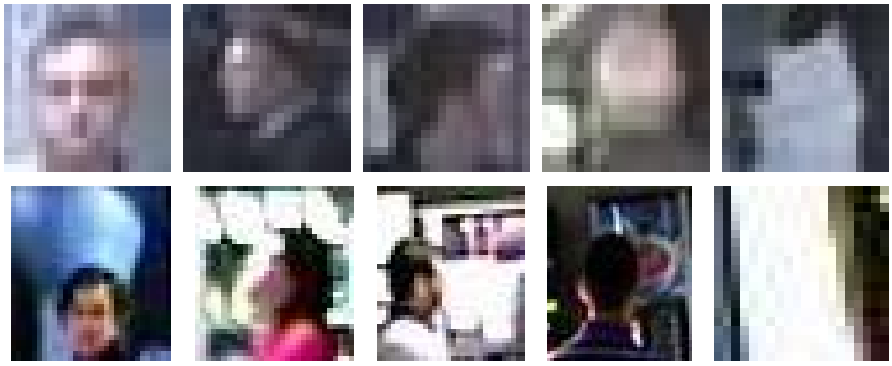
$$\Phi(I,x,y) = \left[X\ Y\ R\ G\ B\ I_x\ I_y\ O\ \mathrm{Gab}_{\{0,\pi/3,\pi/6,4\pi/3\}}\right]. \tag{5}$$

X,Y represent the spatial layout maps in I, and R,G,B are the color channels. I_x and I_y are the directional derivatives of I, and O is the gradient orientation. Finally, Gab is a set of 4 maps containing the results of Gabor filtering, with filters of dimension 2×4, sinusoidal frequency 16, and directions $\mathscr{D} = \{0, \pi/3, \pi/6, 4\pi/3\}$. In order to increase the robustness to local illumination variations, we apply the normalization operator introduced in [77] 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:

$$\widehat{X}_i = \mathrm{diag}(X_I)^{-\frac{1}{2}} X_i \,\mathrm{diag}(X_I)^{-\frac{1}{2}}, \tag{6}$$

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 (see [11] as an example). 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), a multi-class LogitBoost classifier is trained. Each class represent a different head orientation sampled according with a fixed sampling step α and from an extra class containing all the background examples. We experimentally found that $\alpha = 90^\circ$ which correspond to the semantic classes North, South, East and West, is enough for our purposes. Fig. 1 shows some training and testing examples for each class. At testing time, each patch of a sample window (Fig. 2) is independently classified. Then, the classification re-



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

Fig. 1 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. 20×20 pixels).

sult 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, for the sake of brevity).

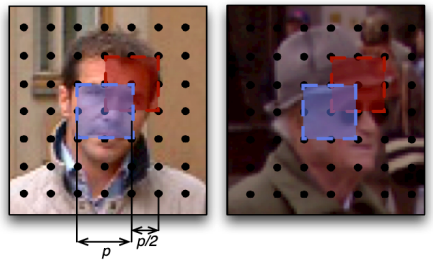


Fig. 2 Array of Covariance matrices (ARCO) feature. The image is organized as a grid of uniformly spaced and overlapping patches. The head orientation result of each patch is estimated by a multi-class classifier.

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 [19], adapted to work on Riemannian manifolds, as suggested by [77, 75]. 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 covari-

ance 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 [77, 75].

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:

$$c = \arg\max_j \{\Delta_j\}, \quad j = 1,\ldots,J. \tag{7}$$

Actually, in our approach, we employ 5 classes mentioned above, i.e., North, South, East, West, and Background. The first four classes indicate the four directions related to the camera orientation. The Background class is introduced to manage cases where 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. We are also aware that the head orientation could be injected in a tracking framework [67], as an additional state that characterizes the human body in this way smoother results should be obtained. However, this smoothing gives poor results and in some cases a drift of the track when dealing with low frame rate videos. Thus, for the sake of the generality, we prefer to keep person tracking and head orientation estimation separated, so as to minimize the error in low frame rate scenario and also they can be used separately in other applications.

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 [77], it is fast, making it suitable for a possible real-time usage.

2.3 Subjective View Frustum Estimation

The Visual Focus Of Attention (VFOA) [74, 40, 67] 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 [1]. 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 [44, 69], and several approximations are considered in many cases. For example, in [74], 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 [67] 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 [40], 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 [31], 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.

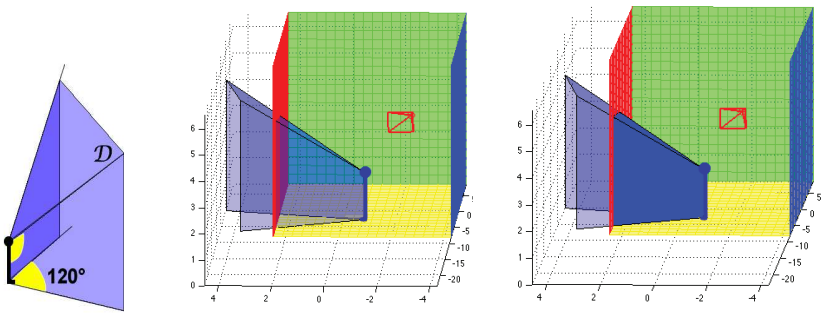


Fig. 3 Left: the SVF model. Center: an example of SVF inside a 3D "box" scene. In red, the surveillance camera position: the SVF orientation is estimated with respect to the principal axes of the camera. Right: the same SVF delimited by the scene constraints (in solid blue).

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 head pose. This assumption has been exploited in several approaches dealing with a meeting scenario [74, 73, 49, 82] or in a smart environment [67, 34]. 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 [16]. This feature represents the three-dimensional (3D) visual field of a human subject in a scene. According to biological evidence [52], the SVF can be modeled as a 3D polyhedron delimiting the portion of the scene that the subject is looking at (see Figure 3).

More in detail, the SVF is defined as the polyhedron $\mathscr{D}$ depicted in Figure 3. 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, while the SVF orientation is obtained by an head pose detector.

The SVF $\mathscr{D}$ 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 (see Figure 3). 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 [57].

3 The Birth of a Group

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 [16], 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 cognitive and social signaling studies, we define the birth of a group when multiple and stable relations are detected over time. In particular, it is highly probable that a relation takes place when two persons are closer than 2 meters [79], and looking at each other [87, 33, 26]. 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 an *Inter-Relation Pattern Matrix* (IRPM), that encodes the social exchanges occurred among all the persons in a scene. The work we present in this section has been published in [4, 15].

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 [46, 43], and person re-identification [91] 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, 9], or smart rooms [83]. 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 (except our first work in [16]), the "subjective" point of view for automated surveillance systems was taken into account by [5], taking inspiration from [60], and it represents therefore the most similar approach in the literature to our work. The difference between [5] and our system are that 1) in [5], 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 signaling 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 [79], 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 Sec. 3.1, the method to build the Inter-Relation Pattern Matrix is described. Then, in Sec. 3.2, experiments and results on home-made and public datasets are shown.

3.1 The Inter-Relation Pattern Matrix

The SVF can be employed as a tool to discover the visual dynamics of the interactions among 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) [79]. 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 [87, 33, 26]. 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 (see Figure 4). The Inter-Relation Pattern Matrix (*IRPM*) records when a possible social interaction occurs, and it can be formalized as a three-dimensional matrix [17], 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.

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 $[t-T+1,t]$. 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 threshold defined *a priori*. This threshold filters

Fig. 4 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 3.1.

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 one-to-one 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. Some examples are depicted in Figures 6, 7 and 8.

3.2 Experimental Results

These experiments aim at showing the capabilities of the proposed approach. We recorded a video sequence of about 3 hours and a half duration, portraying a vending machines area where students take coffee and discuss. 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 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.

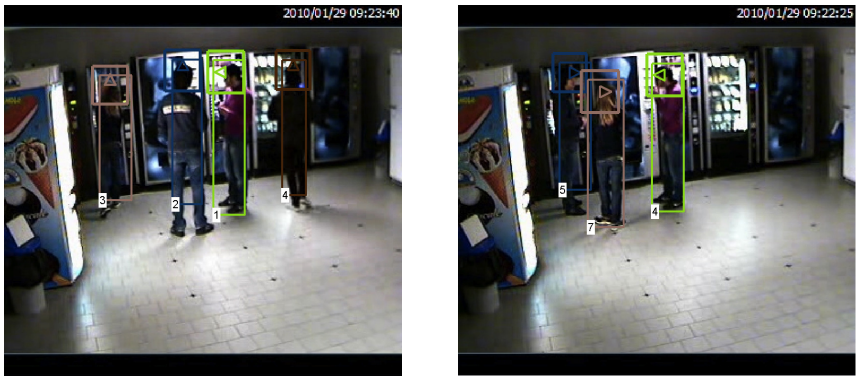


Fig. 5 Examples of tracking and head orientation classification results. The largest box represents the tracking estimation, the smaller box the area where the head is positioned, and the triangle depicts the estimated head orientation.

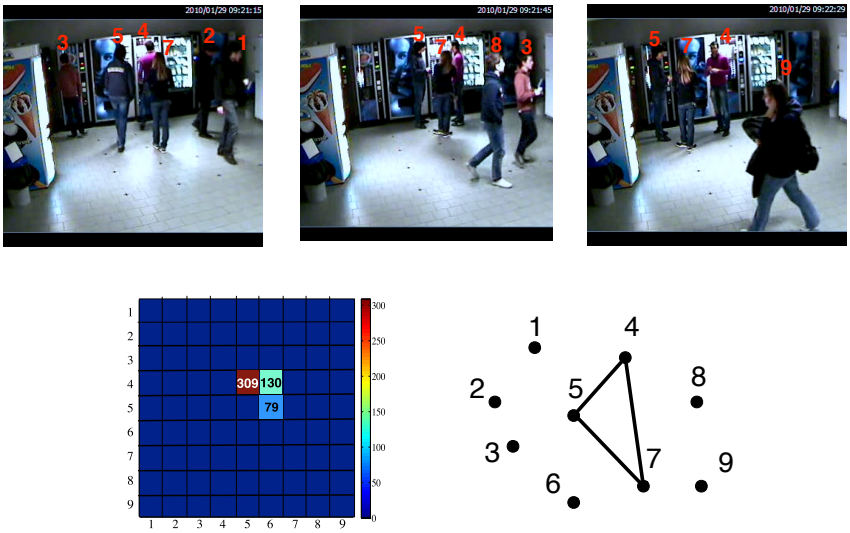


Fig. 6 Example of IRPM analysis of sequence S_{04}. On the top row, some frames of the sequence. On the bottom row, on the left, the cIRPM matrix. Being the cIRPMs symmetric and having null main diagonals, we report for clarity only its strictly upper triangular part. On the right, the corresponding 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[2], 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 (some examples are shown in Figure 5), and construction of the three-dimensional IRPM, indicating which people are potentially interacting at a specific moment.

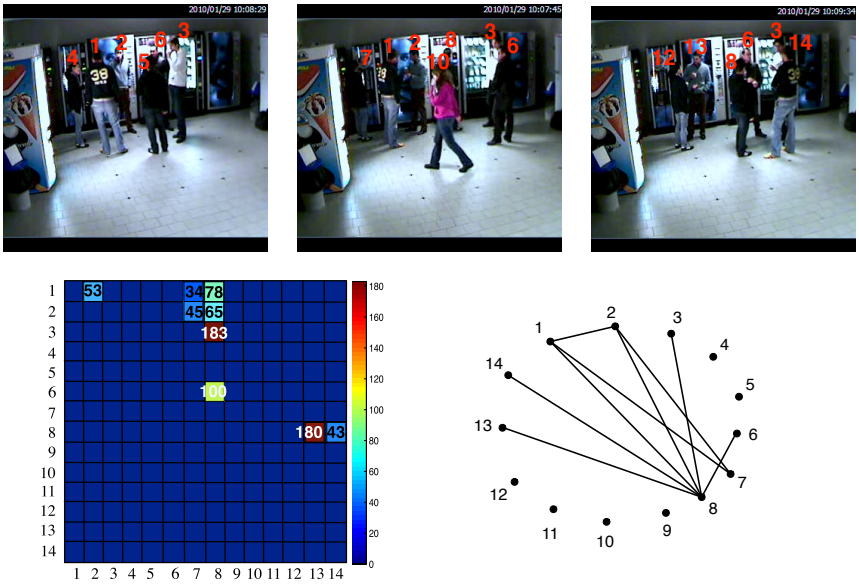


Fig. 7 Example of cIRPM analysis of sequence S_{08}. One big group (1,2,3,6,7,8,13,14) is detected. Note that some people are represented by more than one track, since due to severe or complete occlusions the tracks are sometimes lost and need to be reinitialized (see the text for more details). Person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

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. A qualitative analysis of the results is shown in Figures 6, 7 and 8. 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

[2] The dataset is available at `http://www.lorisbazzani.info/code-datasets/multi-camera-dataset/`

correctly. In particular, Fig. 6 shows the case where a single, small group and other individuals are present in the scene during the recording. In Fig. 7, 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 $(5,8)$. The automatic association between IDs is also possible in such scenarios using re-identification or re-acquisition methods such as [14, 8, 3, 91], but it is out of the scope of this work. Fig. 8 shows that our model is able to detect interactions also when the scene contains multiple groups.

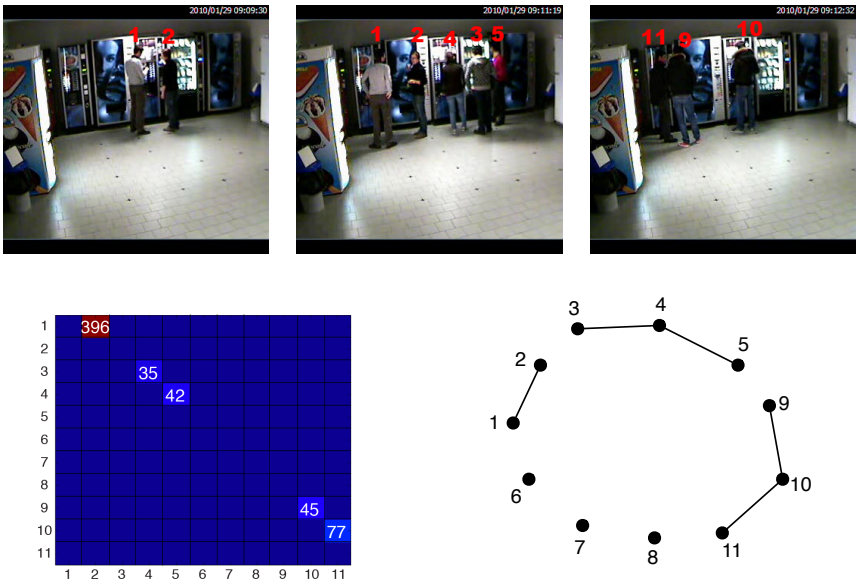


Fig. 8 Example of cIRPM analysis of sequence S_{01}.Three groups (1,2),(3,4,5), and (9,10,11) are detected. One can note that some people are represented by more than one track, since due to severe or complete occlusions the tracks are sometimes lost and need to be reinitialized (e.g. 6,7,8 are reinitialized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

A more sophisticated analysis of accuracy performances of our method is shown in Fig. 9 and 10. 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 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. 9 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 thresh-

old 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. 10 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 shuld 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. 9 and Fig. 10, respectively, decrease before and after the optimal setting of the parameters ($Th = 20$ and $T = 300$).

4 The Evolution of a Group

In this section, we consider the evolution of a group as an instance of tracking, called *group tracking*, published in [2]. Once a group is initialized through cIRPM or employing simple proxemics guidelines, a group can move in the scene, and we propose a tracking approach which embeds the knowledge of the states of the single individuals and the state of the group to provide a robust group localization and tracking.

Group tracking (GT) is of high interest for video surveillance purposes as it allows fine scenario descriptions addressing choral actions that may lead to important threats and highlighting social bounds among individuals. At the same time, it represents a challenging issue, since a group of people is a highly structured entity whose dynamics is complex, and whose appearance is erratic, due to intra- and inter-group occlusions phenomena.

There have been few recent attempts to deal with GT problem. The literature on GT could be partitioned in two sets. The first, *top-down GT*, contains techniques that model groups as blob entities found after background subtraction [46, 85, 43]. In [46], a foreground segmentation method classifies the moving regions in people and groups. In [43], a foreground subtraction-based method models the object paths using a Bayesian network. A set of empirical rules are employed to detect

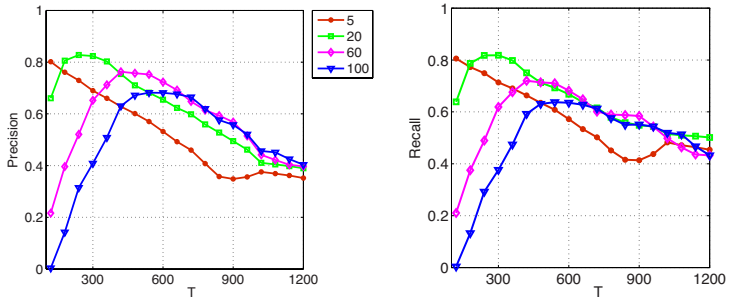


Fig. 9 Evaluation of precision (left) and recall (right) of the proposed method varying the size of the time interval T (x-axis) used to compute the IRPM. The graph shows one curve for each threshold (5, 20, 60 and 100). The maximum for both the statistics is given by setting $Th = 20$.

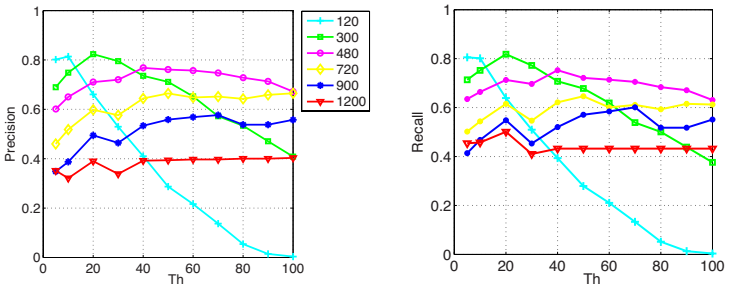


Fig. 10 Evaluation of precision (left) and recall (right) of the proposed method varying the threshold Th (x-axis) used to detect the groups. The graph shows one curve for each time window (120, 300, 480, 720, 900, and 1200). The maximum for both the statistics is given by setting $T = 300$ and the peak corresponds to $Th = 20$.

the groups, however, intra- and inter-group dynamics are not considered in these methods.

The second, *bottom-up GT*, is formed by algorithms that operate after that the individuals have been individually tracked [20, 45, 38, 36]. A set of empirical merging and splitting rules embedded into a Kalman filter are proposed in [20] to track groups. However, the Kalman filter is not able to deal with non-linear dynamics, if not resorting to more complex variants. In [45], a deterministic mass-spring model interprets the result of a multi-object tracker, joining objects sharing a common behavior. In [38], a lattice-based Markov Random Field combined to a particle filter tracks groups as near-regular textures. A method that tracks a group of highly correlated targets by employing a Markov Chain Monte Carlo particle filter is proposed in [36]. However, the last two approaches deal with very constrained intra-group dynamics because they assume a strong correlation among the targets.

In this section, we present a novel way to track groups, namely Collaborative Particle Filters (Co-PF). The underlying idea consists in designing two tracking processes observing a scene under two different perspectives: a low-level, *multi-object*

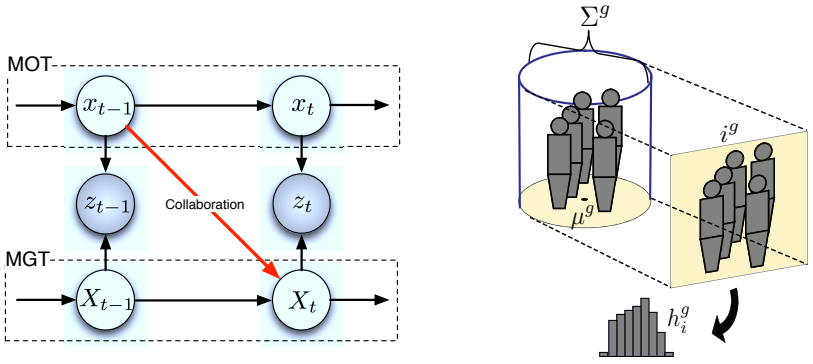
tracker (MOT) performs tracking of multiple individuals, separately; a high-level, *multi-group tracker* (MGT) focuses on groups, and uses the knowledge acquired by the MOT to refine its estimates. Each process consists in a Hybrid-Joint Separable (HJS) filter (see Sec. 2.1), permitting to track multiple entities dealing with occlusions in a very effective way.

The input given by the MOT flows to the MGT in a principled way, *i.e.*, revising the MGT posterior distribution by marginalizing over the MOT's state space. In this way, the MGT posterior is peaked around group configurations formed by trusted individual estimates. In practice, our framework allows to: i) track *multiple* groups; ii) deal with intra- and iii) inter-group occlusions in a 3D calibrated context. The latter two conditions have never been taken into account jointly, and define a brand new operating context, where we put a solid possible solution. Synthetic and real experiments validate our approach and encourage further developments for Co-PF.

The rest of the Section is organized as follows. In Sec.4.1, the collaborative particle filter framework is described and related experiments are presented in Sec.4.2.

4.1 Collaborative Particle Filter

The framework we analyze is sketched in Fig. 11(a): the MOT tracks the individuals in the scene, whereas the MGT tracks groups of individuals. Both the processes share the same observations, $\{z_t\}$, and this highlights our key intuition: the two processes evaluate the scene under two different points of view.



(a) Co-PF (b) Rendering function

Fig. 11 Collaborative PF idea and group rendering.

The MOT process is modeled by HJS filter [35] (Sec. 2.1). Each individual state is modeled as an elliptical shape on the ground plane, *i.e.*, $x^k = \langle \mu^k, \Sigma^k \rangle$, where μ^k is

the position of the individual on the ground plane[3], Σ^k is a covariance that measures the occupancy of the body projected on the ground plane (see [35] for more details).

The MGT process customizes the HJS filter for dealing with groups, and incorporates a fusion component, accepting information from the MOT. We denote the gth group as $X^g = \langle \mu^g, \Sigma^g \rangle$, where μ^g is the 2D position on the floor of the centroid of the gth group and Σ^g is the covariance matrix that approximates the projection of its shape on the floor. 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 [22]. The posterior of the MGT of the g-th group follows the Bayesian recipe (Eq. 1), so that

$$p(X_t^g|z_{1:t}) \propto p(z_t|X_t^g) \int p(X_t^g|X_{t-1}^g)\, p(X_{t-1}^g|z_{1:t-1})\, dX_{t-1}^g. \tag{8}$$

The dynamical model $p(X_t^g|X_{t-1}^g)$ is derived as in Eq. 2, where the joint dynamical model $p(\mathbf{X}_t|\mathbf{X}_{t-1}) \approx p(\mathbf{X}_t) \prod_g q(X_t^g|X_{t-1}^g)$ has $\mathbf{X}_t = \{X_t^1, X_t^2, \ldots, X_t^G\}$, with G the number of groups in the scene. In this case, the function $q(X_t^g|X_{t-1}^g)$ is modeled by considering the nature of $X_t^g = \langle \mu^g, \Sigma^g \rangle$. For the centroid μ^g, we assume a linear motion, perturbed by white noise with parameter σ_μ. The dynamics of the covariance matrix Σ^g is defined by a perturbation of its principal axes, *i.e.*, by varying its eigenvalues $\{\lambda_i\}_{i=1,2}$ and eigenvectors $\{\mathbf{v}_i\}_{i=1,2}$. In particular, we rotate the principal axes by an angle θ, by modifying the eigenvectors: $V' = [R(\mathcal{N}(\theta, \sigma_\theta))\,\mathbf{v}_1, R(\mathcal{N}(\theta, \sigma_\theta))\,\mathbf{v}_2]$ and then, we vary the amplitude of the principal axes by modifying the eigenvalues as follows:

$$\Lambda' = \begin{bmatrix} \mathcal{N}(\lambda_1, \sigma_\lambda) & 0 \\ 0 & \mathcal{N}(\lambda_2, \sigma_\lambda) \end{bmatrix} \tag{9}$$

where $R(\cdot)$ is a rotation matrix and σ_θ and σ_λ are user-defined noise variance values. The matrixes V' and Λ' are then used to recompose the new hypothesis $\Sigma' = V'\Lambda' V'^T$, that will represent a new perturbed elliptical shape. The dynamics prior $p(\mathbf{X}_t^g)$ implements an exclusion principle using Markov Random Fields [35] that cancels out inconsistent hypothesis (*e.g.*, individuals in the same location).

The (single) observation model $p(z_t|X_t^g)$ is derived from Eq.3, where we have $p(z_t|\mathbf{X}_t)$ as joint observation model. In order to easily evaluate an observation z_t, we employ a *rendering function* that maps a state in a convenient feature space[4]. The idea is depicted in Fig. 11(b): when a new group is detected at time t in the scene, its centroid μ^g and occupancy area Σ^g are robustly estimated, forming the initial state X_t^g. The rendering function builds a volume of height 1.80m upon the area Σ^g, in order to surround the people of the group. From this volume, the projection i^g (namely, the model of X_t^g) on the image plane is evaluated, and finally, the histogram h_i^g is computed. This function permits to estimate novel state hypotheses X'^g_t: given its components $\langle \mu'^g, \Sigma'^g \rangle$, the rendering function takes the model i^g

[3] Please note that the ground plane position is inferred employing the calibration of the camera.

[4] This is analogue to what was done in [35] for the single individuals.

deforming it opportunely (by a re-scaling, considering the μ'^g, and by a shearing, taking into account the deformation resulted by the perturbation of the covariance matrix Σ'^g). This brings to a novel h'^g_i, which is compared with the observation estimated directly from the scene by the rendering function applied to $\langle \mu'^g, \Sigma'^g \rangle$. We use the Bhattacharyya distance as similarity measurement.

The joint observation model $p(z_t|\mathbf{X}_t)$ mirrors what part of the group X^g_t is visible (not occluded) by taking into account the remaining groups $\mathbf{X}^{\neg g}_t$. This encodes at the same time the advantages and limitations of the observation model. Actually, we assume a group as a rigid solid shape (the model i^g), 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.

Co-PF solves this problem, and permits a very fine estimation of the whereabouts of a scene, making the group tracking very robust. It basically injects the information collected by the MOT into the MGT. Considering the filtering expression in Eq. 8, the fusion occurs on the posterior at time $t-1$:

$$p(X^g_{t-1}|z_{1:t-1}) \propto \int p(X^g_{t-1}|\mathbf{x}_{t-1}, z_{1:t-1})\, p(\mathbf{x}_{t-1}|z_{1:t-1})\, d\mathbf{x}_{t-1} \tag{10}$$

The first term of Eq. 10 is the core of our approach as it revises the group posterior distribution at time $t-1$, also considering the states of the single individuals. In this way, the second term (the posterior at time $t-1$ of the MOT process) 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) [62], that decomposes a structured conditioned distribution as a convex combination of pairwise conditioned distributions. This leads to:

$$p(X^g_{t-1}|\mathbf{x}_{t-1}, z_{1:t-1}) \approx \alpha_1 p(X^g_{t-1}|\mathbf{x}_{t-1}) + \alpha_2 p(X^g_{t-1}|z_{1:t-1}), \tag{11}$$

where $\alpha_1, \alpha_2 > 0$ and $\alpha_1 + \alpha_2 = 1$. We can now rewrite Eq. 10 as:

$$p(X^g_{t-1}|z_{1:t-1}) \approx \alpha_1 \int p(\mathbf{x}_{t-1}|z_{1:t-1})\, p(X^g_{t-1}|\mathbf{x}_{t-1})\, d\mathbf{x}_{t-1} + \tag{12}$$

$$\alpha_2\, p(X^g_{t-1}|z_{1:t-1}) \underbrace{\int p(\mathbf{x}_{t-1}|z_{1:t-1})\, d\mathbf{x}_{t-1}}_{=1}. \tag{13}$$

At this point, it is easy to realize that $p(X^g_{t-1}|z_{1:t-1})$ becomes a combination of the natural group posterior and a marginalization of the *linking* probability $p(X^g_{t-1}|\mathbf{x}_{t-1})$, that relates a group to individuals, weighted by the MOT posterior. In other words, the group posterior is revisited by injecting in a principled way the information on the single targets (the MOT posterior), conveyed selectively by $p(X^g_{t-1}|\mathbf{x}_{t-1})$. An example will demonstrate the advantage of this formulation.

The linking probability $p(X^g_{t-1}|\mathbf{x}_{t-1})$ is factorized as an MMP as follows:

$$p(X^g_{t-1}|\mathbf{x}_{t-1}) \approx \sum_{k=1}^{K} p(X^g_{t-1}|x^k_{t-1})\,\beta^{k,g} \tag{14}$$

$$\propto \sum_{k=1}^{K} p(x^k_{t-1}|X^g_{t-1})\,p(X^g_{t-1})\,\beta^{k,g} \tag{15}$$

where $\beta^{k,g} > 0\ \forall k,g$ and $\sum_k \beta^{k,g} = 1$. Each term of the sum in Eq. 14 represents the posterior probability that the gth group X^g_{t-1} contains the kth target x^k_{t-1}.

In Eq. 15, the posterior is modeled employing the Bayes rule, where $p(x^k_{t-1}|X^g_{t-1})$ defines the *linking* likelihood that each single individual state x^k_{t-1} is a subpart of X^g_{t-1}. Hence, we define a probability model based on three components: 1) appearance similarity, 2) dynamics consistency, and 3) group membership. The appearance similarity is encoded by the Bhattacharyya distance between the HSV histograms of the two entities: $d_{\mathrm{HSV}}(X^g_{t-1},x^k_{t-1})$. 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 $d_{\mathrm{dir}}(X^g_{t-1},x^k_{t-1}) = |1-|\mathrm{dir}(X^g_{t-1}) - \mathrm{dir}(x^k_{t-1})|/\pi|$, 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:

$$d_{\mathrm{mbr}}(X^g_{t-1},x^k_{t-1}) = \begin{cases} 1 \text{ if } & x^k_{t-1} \in X^g_{t-1} \\ 0 \text{ otherwise} & \end{cases} \tag{16}$$

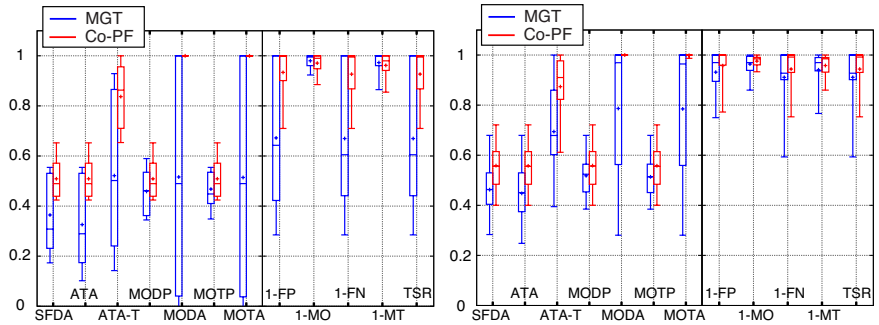
where the membership operator $\in$ controls if the kth person position is inside the gth group ellipse. Therefore, $p(x^k_{t-1}|X^g_{t-1}) = d_{\mathrm{HSV}}(X^g_{t-1},x^k_{t-1}) \cdot d_{\mathrm{dir}}(X^g_{t-1},x^k_{t-1}) \cdot d_{\mathrm{mbr}}(X^g_{t-1},x^k_{t-1})$. The coefficients $\beta^{k,g}$ express a linking preference that an object belongs to a group, and are left here as uniform, *i.e.*, $\beta^{k,g} = 1/G$.

Finally, the prior $p(X^g_{t-1})$ discards the biggest and the smallest group hypotheses, rejecting the particles in which the size of the group is below a threshold τ_b or above a threshold τ_a.

An example that explains the strength of our formulation can be represented by an intra-group occlusion in the gth group at time $t-1$, which is very common due to the dynamical nature of a group of moving people. Let x^k_{t-1} a target of the group X^g_{t-1} that vanishes as occluded by the remaining individuals of that group. The group posterior $p(X^g_{t-1}|z_{1:t-1})$ will not be very high, for the limits of the visual, rigid, group representation. However, the MOT process, dealing with single objects and managing their occlusions, will "understand" the fact that x^k_{t-1} is occluded, producing a high $p(x^k_{t-1}|z_{1:t-1})$. This probability value will flow through $p(X^g_{t-1}|\mathbf{x}_{t-1})$, which is high because, even if occluded, the position and the velocity of x^k_{t-1} are correctly estimated by the MOT process, and will give a high linking likelihood. This will reinforce the final estimation of the hybrid posterior for X^g_{t-1}, thus permitting to estimate the subsequent group sample set in a more correct way.

4.2 Experimental Results

Our approach has been evaluated on synthetic data and publicly available datasets (PETS 2006[5] and PETS 2009[6]). We carried out a comparative analysis with respect to the MGT (without the proposed collaboration stage), highlighting that Co-PF is more able 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.



(a) PETS 2006 results (b) PETS 2009 results

Fig. 12 Statistics on the synthetic test set.

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 PETS dataset 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 follow measures: Average Tracking Accuracy (ATA), Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), False Positive (FP), Multiple Objects

[5] `http://www.cvg.rdg.ac.uk/PETS2006/`

[6] `http://www.cvg.rdg.ac.uk/PETS2009/`

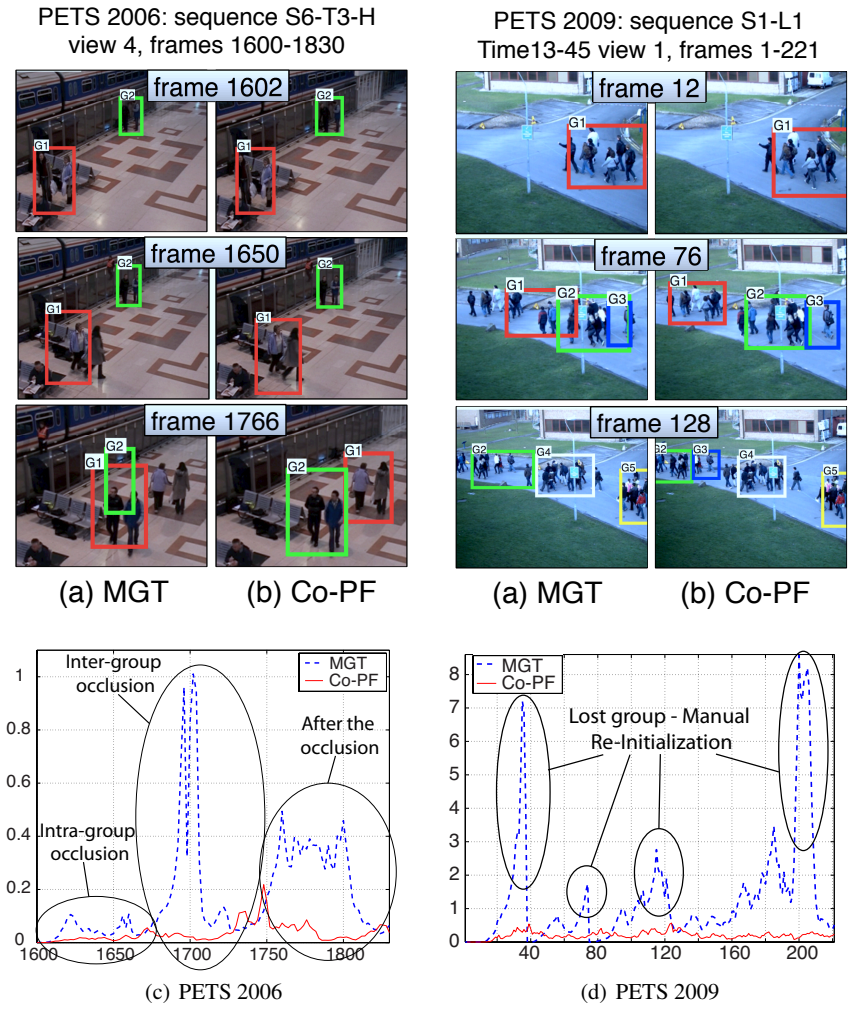


Fig. 13 Comparison of MGT (first and third column) and Co-PF (second and fourth column) on PETS 2006 and PETS 2009. The second row compares the PF uncertainty [42] in the two experiments.

(MO), False Negative (FN), Tracking Success Rate (TSR) and so on (further details in [29, 68]). For each measure, a *boxplot* representation is given [29], 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 (Fig. 12(a) and Fig. 12(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 μ^g and Σ^g to the projections of the individuals new entries on the ground plane. If lost, a group is manually reinitialized. Note that group split and merge is not modeled in this probabilistic framework. It is an hard problem that has to be handled as future work. We show here two representative examples. In real scenarios, MGT is not able to deal completely with the intra- and inter-group dynamics (Fig. 13(a)). On the other hand, Co-PF exploits the MOT results, enriching the posterior knowledge given by the MGT (Fig. 13(b)).

To give further support to our Co-PF, we evaluate the uncertainty of the particle filters [42]. Fig. 13(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). Fig. 13(d) shows a similar behavior of Fig. 13(c), highlighting that the MGT looses the tracks several times.

5 Human-Environment Interactions: Interest Maps

The contribution of this Section is a visualization application of the SFV-based framework (see Sec. 2.3), called the *Interest Map*, published in [16]. Since the part of a scene that intersects the SVF is the area observed by the SVF owner, we collect this information for each subject, over a given time interval. This permits to infer which are the parts of the scene that are more observed, 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). 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 5.1 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 5.2). The interest map generation process is presented in Section 5.3, and experiments, reported in Section 5.4, show qualitative results of the interest map given the monitored environment.

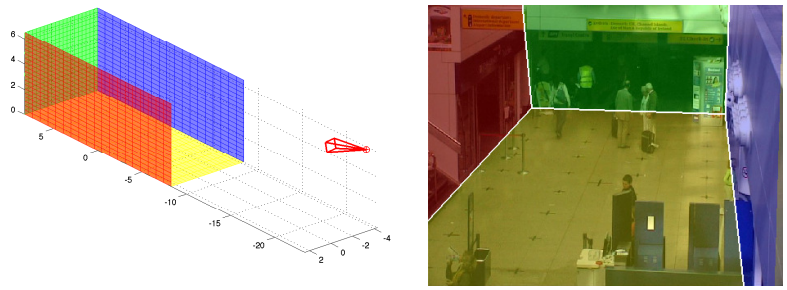
5.1 3D Map Estimation

Let's 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 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 P is given by two equations:

$$u = \frac{\mathbf{p}_1^\top \mathbf{M}}{\mathbf{p}_3^\top \mathbf{M}}, \quad v = \frac{\mathbf{p}_2^\top \mathbf{M}}{\mathbf{p}_3^\top \mathbf{M}}, \quad \text{with } \mathsf{P} = \begin{bmatrix} \mathbf{p}_1^\top \\ \mathbf{p}_2^\top \\ \mathbf{p}_3^\top \end{bmatrix}. \tag{17}$$

(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 14). 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 [70, 21, 7].



(a) (b)

Fig. 14 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.

5.2 Motion Orientation Estimation

The tracking algorithm of Sec. 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, this approach can be seen as an alternative, yet simpler solution to the method proposed in Sec. 2.2, 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.

5.3 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 $\mathscr{D}_{i,t}$ 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 Sec. 5.1. A full volumetric reasoning could be considered too, but this would capture other kinds of information, such as people interactions.

Each SVF $\mathscr{D}_{i,t}$ at current time is projected on each scene plane. This is equivalent to estimate the vertices of $\mathscr{D}_{i,t}$ 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 the image plane. Two examples of the projected SVF are shown in Figure 15. 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 $T - \tau$ to T, where τ 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 τ is the chosen size of the time window.

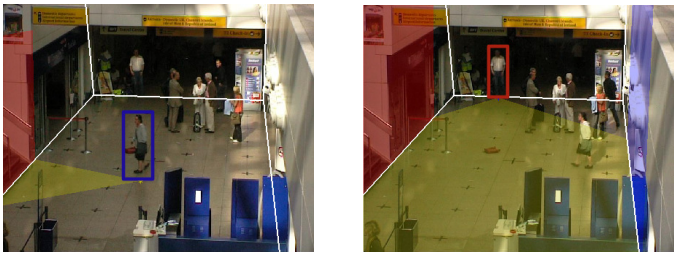


Fig. 15 Two examples of projection of the SVF on the scene's main planes. The 3D map permits to suitably model the interactions of the SVF with the scene.

5.4 Experimental Results

We perform some tests over the publicly available PETS 2007 sequence sets[7], aiming at showing the expressiveness of our framework on widely known and used datasets. Two sequences are taken into account for the experimental validation, both belong to the S07 dataset depiciting an airport area monitoring. The first sequence is captured by Camera 2, the second one is captured by Camera 4.

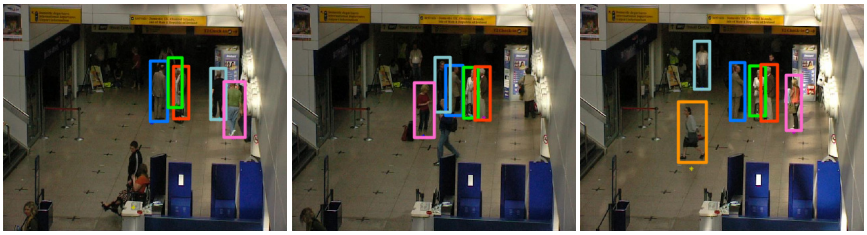
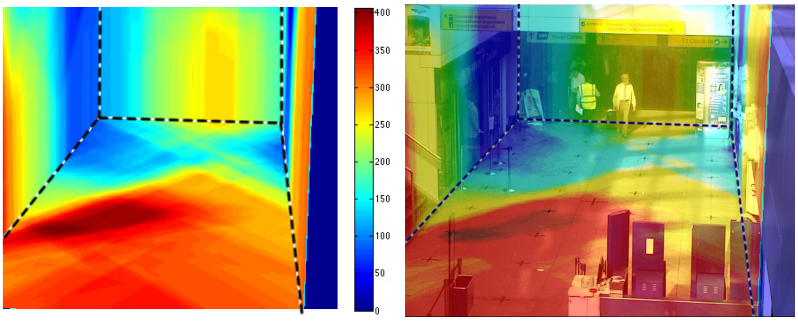


Fig. 16 Some frames of the sequence from camera 2. The bounding boxes highlight the tracking results.

In Figure 16 we 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 Figure 17, 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 18. In this case, the most seen areas of the parallelepiped (the 3D map)

[7] `http://www.cvg.rdg.ac.uk/PETS2007/`



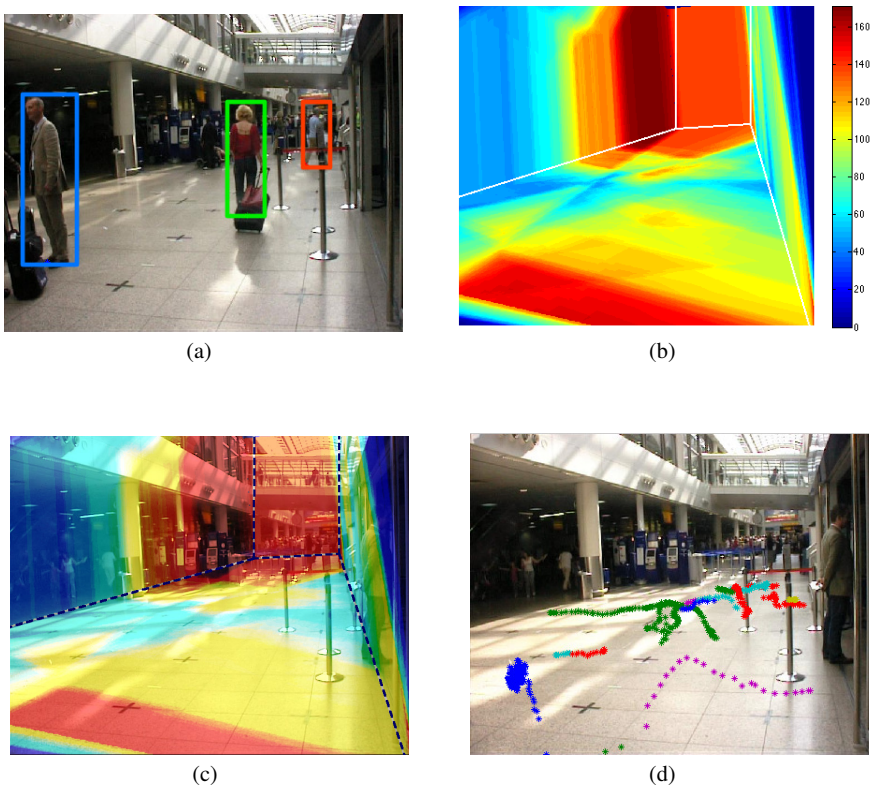
(a) (b)

Fig. 17 (a) Interest Map for S07 sequence from camera 2 ("hot" colors represent the areas more frequently observed, and the opposite for the "cold" areas). (b) The same Interest Map superimposed on one frame of the sequence.

are two (Fig. 18(b)-(c)). 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 (Fig. 18(d)) the tracking results. This representation is less meaningful from the point of view of the analysis of the people attention. Our information visualization technique is instead intuitive and it captures in a very simple and richer way where people attention is focused.

6 Conclusions

This chapter presents a set of techniques for managing groups and group activities, 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 abstract 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 is one of the early example of how computer vision and social signaling may collaborate for a new level of the video surveillance research.



(a) (b)

(c) (d)

Fig. 18 (a) One frame of sequence S07 camera 4, with the tracking results. (b) The obtained Interest Map ("hot" colors represent the areas more frequently observed, and the opposite for the "cold" areas). (c) The same Interest Map superimposed on one frame. (d) The tracks of the 4 people estimated throughout the sequence displayed in the same frame.

References

[1] 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.

[2] L. Bazzani, M. Cristani, and V. Murino. Collaborative particle filters for group tracking. In *IEEE International Conference on Image Processing*, 2010.

[3] 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.

[4] 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*, 2011. in print.

[5] B. Benfold and I. Reid. Guiding visual surveillance by tracking human attention. In *Proceedings of the 20th British Machine Vision Conference*, September

2009.

[6] L. Breiman, JH Friedman, R. Olshen, and CJ Stone. Classification and Regression Trees. *Ann. Math. Statist.*, 19:293–325, 1984.

[7] 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.

[8] 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. in print.

[9] T. Choudhury and A. Pentland. The sociometer: A wearable device for understanding human networks. In *CSCW - Workshop on ACCUCE*, 2002.

[10] 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.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[12] Arnaud Doucet, Nando De Freitas, and Neil Gordon, editors. *Sequential Monte Carlo methods in practice*. Springer, 2001.

[13] P. Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384, 1993.

[14] 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.

[15] 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.

[16] Michela Farenzena, Loris Bazzani, Vittorio Murino, and Marco Cristani. Towards a subject-centered analysis for automated video surveillance. In *Proceedings of the 15th International Conference on Image Analysis and Processing*, ICIAP '09, pages 481–489, Berlin, Heidelberg, 2009. Springer-Verlag.

[17] L. Freeman. Social networks and the structure experiment. In *Research Methods in Social Network Analysis*, pages 11–40, 1989.

[18] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[19] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *The annals of statistics*, 28(2):337–374, 2000.

[20] G. Gennari and G. D. Hager. Probabilistic data association methods in visual tracking of groups. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[21] 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.

[22] E.T. Hall. *The hidden dimension*, volume 6. Doubleday New York, 1966.

[23] S. Hongeng and R. Nevatia. Large-scale event detection using semi-hidden markov models. In *IEEE International Conference on Computer Vision*, volume 2, 2003.

[24] M Isard and A Blake. Condensation: Conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29:5–28, 1998.

[25] M. Isard and J. MacCormick. BraMBLe: a bayesian multiple-blob tracker. In *Int. Conf Computer Vision*, volume 2, pages 34–41, 2001.

[26] 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.

[27] 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.

[28] R E Kalman. A new approach to linear filtering and prediction problems. *Tran. of the ASME Journal of Basic Engineering*, (82):35–45, 1960.

[29] 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.

[30] M.L. Knapp and J.A. Hall. *Nonverbal communication in human interaction*. Wadsworth Pub Co, 2009.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] Oswald Lanz. Approximate bayesian multibody tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28:1436–1449, September 2006.

[36] Y. Lao and F. Zheng. Tracking a group of highly correlated targets. In *IEEE International Conference on Image Processing*, 2009.

[37] 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.

[38] W.C. Lin and Y. Liu. A lattice-based mrf model for dynamic near-regular texture tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(5):777–792, 2007.

[39] 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.

[40] 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.

[41] 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.

[42] E. Maggio, F. Smerladi, and A. Cavallaro. Combining colour and orientation for adaptive particle filter-based tracking. In *British Machine Vision Conference*, 2005.

[43] Jorge S Marques, Pedro M Jorge, Arnaldo J Abrantes, and J M Lemos. Tracking groups of pedestrians in video sequences. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop*, volume 9, pages 101–101, Jun 2003.

[44] 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.

[45] T. Mauthner, M. Donoser, and H. Bischof. Robust tracking of spatial related components. In *IEEE International Conference on Pattern Recognition*, pages 1–4, Dec. 2008.

[46] Stephen J. Mckenna, Sumer Jabri, Zoran Duric, Harry Wechsler, and Azriel Rosenfeld. Tracking groups of people. *Computer Vision and Image Understanding*, 2000.

[47] 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.

[48] Bingbing Ni, Shuicheng Yan, and Ashraf A. Kassim. Recognizing human group activities with localized causalities. In *CVPR'09*, pages 1470–1477, 2009.

[49] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase. Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 1175–1180, New York, NY, USA, 2006. ACM.

[50] 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.

[51] P. Pan and D. Schonfeld. Dynamic proposal variance and optimal particle allocation in particle filtering for video tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(9):1268–1279, 2008.

[52] J. Panero and M. Zelnik. *Human dimension & interior space: a source book of design reference standards*. Whitney Library of Design, 1979.

[53] 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.

[54] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: modeling social behavior for multi-target tracking. In *Proc. 12th International Conference on Computer Vision, Kyoto, Japan*, Proc. 12th International Conference on Computer Vision, Kyoto, Japan, 2009.

[55] A. Pentland and S. Pentland. *Honest signals: how they shape our world*. The MIT Press, 2008.

[56] Alex Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:107–119, January 2000.

[57] F.P. Preparata and M.I. Shamos. *Computational geometry: an introduction*. Springer, 1985.

[58] G. Psathas. *Conversation analysis: The study of talk-in-interaction*. Sage Publications, Inc, 1995.

[59] V.P. Richmond, J.C. McCroskey, and S.K. Payne. *Nonverbal behavior in interpersonal relations*. Allyn and Bacon, 2000.

[60] Neil Robertson and Ian Reid. *Estimating Gaze Direction from Low-Resolution Faces in Video*. 2006.

[61] R.J. Rummel. *Understanding conflict and war*. Sage Publications, 1981.

[62] L.K. Saul and M.I. Jordan. Mixed memory markov models: Decomposing complex stochastic processes as mixtures of simpler ones. *Machine Learning*, 37(1):75–87, 1999.

[63] R.E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Mach. Learn.*, 37:297–336, 1999.

[64] AE Scheflen. The significance of posture in communication systems. *Communication Theory*, page 293, 2007.

[65] K.R. Scherer. *Personality markers in speech*. Cambridge Univ. Press, 1979.

[66] P. Scovanner and M.F. Tappen. Learning pedestrian dynamics from the real world. In *IEEE International Conference on Computer Vision*, pages 381–388, 2009.

[67] K. Smith, S. Ba, J. 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):1–18, 2008.

[68] K. Smith, D. Gatica-Perez, J. Odobez, and S. Ba. Evaluating multi-object tracking. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 36–43, 2005.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] M. Viola, M.J. Jones, and P. Viola. Fast multi-view face detection. In *Proc. of Computer Vision and Pattern Recognition*. Citeseer, 2003.

[82] 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.

[83] 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.

[84] 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.

[85] 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. In *IEEE International Conference on Pattern Recognition*, 2006.

[86] 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.

[87] 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.

[88] 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.

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

[90] 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.

[91] W. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *Proceedings of the British Machine Vision Conference*, 2009.