J Real-Time Image Proc
DOI 10.1007/s11554-014-0428-8

SPECIAL ISSUE PAPER

Predicting and recognizing human interactions in public spaces

Fabio Poiesi · Andrea Cavallaro

Received: 14 September 2013 / Accepted: 25 April 2014

Abstract We present an extensive survey of methods for recognizing human interactions and propose a method for predicting rendezvous areas in observable and unobservable regions using sparse motion information. Rendezvous areas indicate where people are likely to interact with each other or with static objects (e.g., a door, an information desk or a meeting point). The proposed method infers the direction of movement by calculating prediction lines from displacement vectors and temporally accumulates intersecting locations generated by prediction lines. The intersections are then used as candidate rendezvous areas and modeled as spatial probability density functions using Gaussian Mixture Models. We validate the proposed method to predict dynamic and static rendezvous areas on real-world datasets and compare it with related approaches.

Keywords Human interactions · Motion prediction · Behavior understanding · Crowd analysis · Gaussian mixture models

This work was supported in part by the Artemis JU and in part by the UK Technology Strategy Board through COPCAMS Project under Grant 332913.

F. Poiesi (✉) · A. Cavallaro
Centre for Intelligent Sensing, Queen Mary University of London, Mile End Road, London E1 4NS, UK
e-mail: fabio.poiesi@qmul.ac.uk

A. Cavallaro
e-mail: a.cavallaro@qmul.ac.uk

1 Introduction

An interaction occurs "*when two or more people or things communicate with or react to each other*".[1] We consider an interaction to involve people reacting to each other and coordinating their behavior (movements) with respect to a common interest, for example a meeting point or an incident. Interactions among two or more people (or between a person and an object) are characterized by *where* and *when*. Neither the exact location nor the time are specified in the above definition. In fact, a reaction among people occurs only when specific movements and spatio-temporal conditions are simultaneously acknowledged. Therefore, the concept of *when* is a direct consequence of *where*, and vice versa. In order to recognize such interactions, we need to describe a meaningful set of features that can highlight coordinated behaviors and validate them over time.

The existence of a common interest among people in a scene can be defined based on relationships among people and rendezvous areas in (or just outside) the image. When two people are approaching a common location the rendezvous area will be the location where these people are likely to interact (Fig. 1). The meaning of rendezvous areas is supported by the fourth law of Gestalt, the *law of common fate* [36]. This law refers to the directional lines that people perceive through a design or layout. These lines are not explicitly represented, but are perceived from the dynamics of the design. If two people are moving towards the same destination, based on their pose it is possible to create a virtual directional line, which is known as the law of common fate, and their pose can be used to infer their future path [67]. This concept is also employed by painters

[1] Definition taken from Cambridge Dictionary, Cambridge University Press 2014.

Published online: 15 May 2014

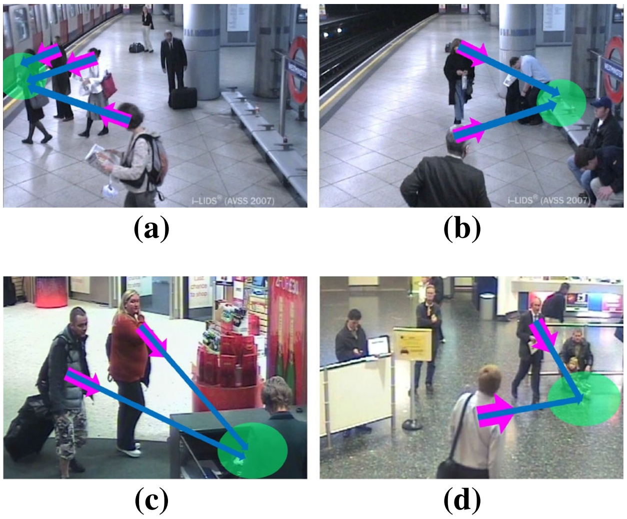


Fig. 1 Conceptual representation of the assumptions used for the localization of rendezvous areas. In two different scenarios the same assumptions can be used to localize different centers of attention. **a** People that are meeting to get on the train, **b** two people meeting to sit on the same bench, **c** two people with the same direction towards the information desk and **d** two people that are meeting at the arrivals of the airport

when they depict a scenario with two important objects aiming at or moving towards a common point of interest.

Rendezvous areas can be static or dynamic. When the movement of objects shows common interest with respect to points, the agreement about their importance is globally verified among the involved objects. If movements are involved, a possible way to understand whether points are of attention is by considering them as attractors for the interested objects. The surrounding objects keep spatio-temporal relationships with the points and, for this reason, the importance of these points may be determined when the objects attempt to spatially interact with them. These regions of interest can be used to predict whether people are moving to forbidden areas, to study evacuation plans or to automatically locate areas where advertisements or passenger information can be displayed.

In this paper, we survey state-of-the-art methods for the recognition of human interactions. Moreover, we propose a method for the prediction of (dynamic and static) rendezvous areas that uses frame-by-frame object motion to infer prediction lines to determine candidate convergence locations [53]. Dynamic rendezvous areas require a short buffer to collect candidate convergence locations. Static rendezvous areas are calculated offline by spatio-temporal integration of the dynamic rendezvous areas. Smooth and continuous prediction of rendezvous areas are inferred by fitting Gaussian Mixture Models (GMM) on candidate convergence locations. Both object motion estimation and GMMs can be implemented in real time [22, 58].

The paper is organized as follows. Section 2 describes state-of-the-art methods for the recognition of human interactions. Section 3 introduces and motivates the problem of predicting human interactions. Section 4 presents the proposed approach for predicting human interactions and in Sect. 6 we validate the method. Section 7 draws the conclusions and discusses future research directions.

2 Interaction analysis: taxonomy and survey

In this section, we describe state-of-the-art methods to recognize human interactions. Figure 2 summarizes the methods presented in this section.

2.1 Global vs. individual motion

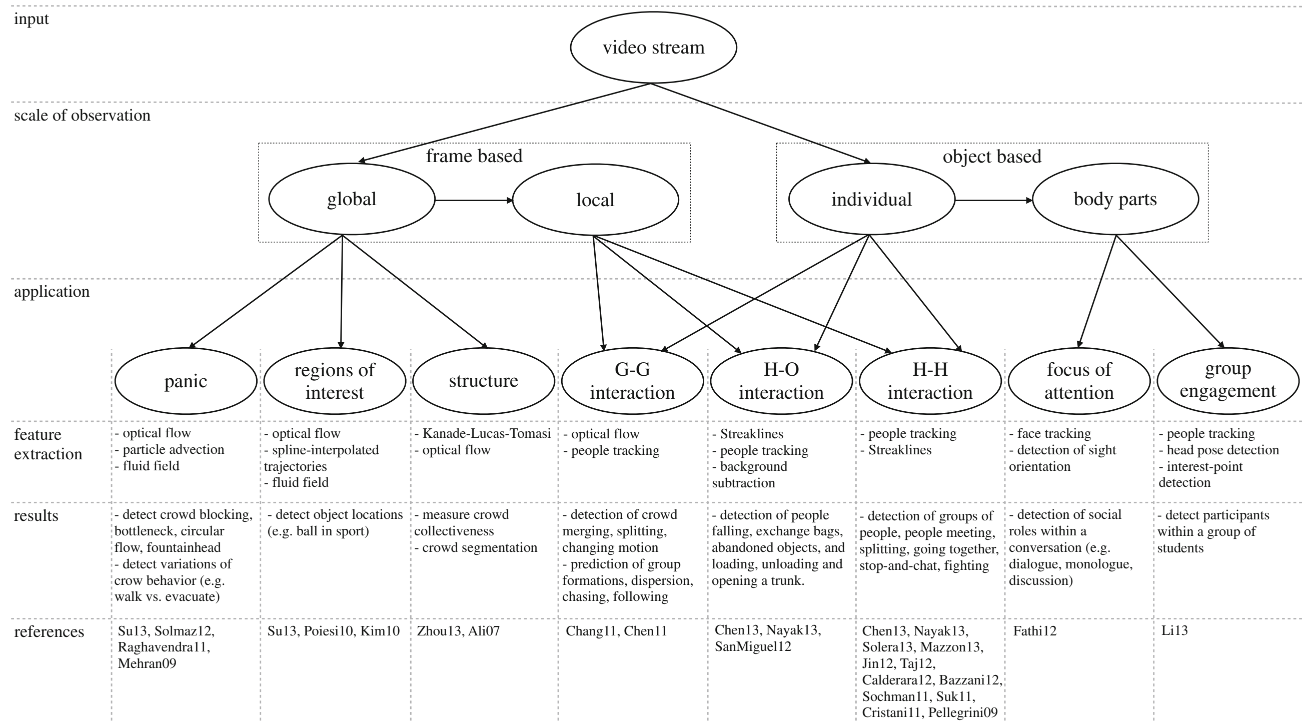
The recognition of interactions in video streams is highly dependent on the scale of observation (e.g., global or individual).

Global refers to methods that infer interactions not relying on descriptors of independent entities in the scene (e.g., objects, people), but characterizing over time salient scene elements (e.g., interest points, moving patches). These methods are usually applied in the case of high-density crowds or challenging view-points (e.g., with multiple occlusions). These methods can be directly used for detecting global interactions such as panic situations [61], or can be clustered to detect more specific (*local*) interactions such as a fight [47]. Global motion patterns are usually computed at frame level based on variations of intensities (or their derivatives) between a few consecutive frames.

Individual refers to methods that infer interactions relying on motion patterns extracted for each object in the scene (e.g., trajectories). Individual motion patterns use object models that are matched frame-by-frame to find correspondences over long periods of time. Individual motion patterns can be directly used for detecting interactions such as one-to-one interactions, groupings and group interactions [5, 43], or to detect *body parts* to infer detailed interactions, such as the engagement of individuals within a group (e.g., conversations) [37].

2.2 Frame-based methods

Frame-based methods can be based on global or local patterns. Global patterns directly employ the estimated motion calculated in a frame to infer interaction properties. Methods based on local patterns process the estimated motion in order to reduce it into clusters each containing uniform motion characteristics (e.g., direction or magnitude) before inferring interaction properties.



	panic	regions of interest	structure	G-G interaction	H-O interaction	H-H interaction	focus of attention	group engagement
feature extraction	- optical flow - particle advection - fluid field	- optical flow - spline-interpolated trajectories - fluid field	- Kanade-Lucas-Tomasi - optical flow	- optical flow - people tracking	- Streaklines - people tracking - background subtraction	- people tracking - Streaklines	- face tracking - detection of sight orientation	- people tracking - head pose detection - interest-point detection
results	- detect crowd blocking, bottleneck, circular flow, fountainhead - detect variations of crow behavior (e.g. walk vs. evacuate)	- detect object locations (e.g. ball in sport)	- measure crowd collectiveness - crowd segmentation	- detection of crowd merging, splitting, changing motion - prediction of group formations, dispersion, chasing, following	- detection of people falling, exchange bags, abandoned objects, and loading, unloading and opening a trunk.	- detection of groups of people, people meeting, splitting, going together, stop-and-chat, fighting	- detection of social roles within a conversation (e.g. dialogue, monologue, discussion)	- detect participants within a group of students
references	Su13, Solmaz12, Raghavendra11, Mehran09	Su13, Poiesi10, Kim10	Zhou13, Ali07	Chang11, Chen11	Chen13, Nayak13, SanMiguel12	Chen13, Nayak13, Solera13, Mazzon13, Jin12, Taj12, Calderara12, Bazzani12, Sochman11, Suk11, Cristani11, Pellegrini09	Fathi12	Li13

Fig. 2 Summary of interaction recognition approaches. Each approach includes information about feature extraction methods, examples of results and references. Key *G–G* Group–Group, *H–O* Human–Object, *H–H* Human–Human

2.2.1 Global patterns

Motion patterns can be globally extracted using techniques based on optical flow [42] or with single-frame descriptors extracted within spatio-temporal volumes such as cuboids [34].

Using optical flow, motion patterns can be extracted for specific intervals of time by using the particle advection approach [45]. Particle advection involves a grid of particles on the image plane. Each particle accumulates the optical flow in its neighborhood to characterize local motion to enable the characterization of panic situations, such as blocking (i.e., people moving in different directions blocking each other), lanes (i.e., large flows of people moving in opposite directions), bottleneck (i.e., people converging on common (narrow) locations), ring (i.e., coordinated circular flow of a crowd) or fountainhead (i.e., people diverging from a location) [61].

An alternative to particle advection is modeling crowd motion as a viscous fluid followed by an analysis based on the combination of a spatio-temporal variation fluid field (i.e., local pixel fluctuations) and a spatio-temporal viscous force field (i.e., relationships among different local fluctuations) [62]. The degree of individuals acting as a union within dense moving crowds can be studied by quantifying collectiveness [72]. Using optical flow, it is also possible to automatically detect reoccurring motion patterns such as congestion situations, which are due to the lateral oscillations of people's torsos [31]. These oscillations are the result of low velocities of people in congested crowds and can be identified by computing symmetry measures from histograms of optical flow vectors.

Interacting people can act as a whole when they have a common interest. For example, a ball in a basketball match (or soccer) can be assumed to be the object that most of the players are interacting with and by extracting the regions of motion convergence of the players, it is possible to predict the location of the ball [29, 53]. Specifically, regions of convergence can be either obtained by extending the direction of the motion vectors and calculating their intersections [53], or by computing the convergence on a dense motion field, which can in turn be obtained by applying a spline-interpolation method on player's tracks [29]. Interactions are also analyzed to detect abnormal events, such as panic situations [13, 45, 54]. When the particle advection method is used [45], in the accumulation point of each particle, the Social Force Model (SFM) [25] can be applied and interaction forces can be analyzed to detect panic behaviors generated by sudden variations in human interactions.

2.2.2 Local patterns

The global motion can be spatio-temporally clustered (or segmented) in order to infer more specific interactions. For example, Streaklines, a clustering method for flow fields, can be integrated with particle advection for a more accurate analysis of spatial changes in the particle flow to capture sudden motion variations [46]. After applying the SFM to the Streaklines, a set of features learned with Support Vector Machine (SVM) can be employed to discriminate between normal and abnormal behaviors, and to localize convergence and divergence regions of motion in panic situations. Moreover, Streaklines enable the analysis of two-person interactions (e.g., fighting, shaking hands) or between individuals and objects (e.g., a person loading a trunk or entering in a vehicle) [47]. Such analysis is achieved by temporally segmenting Streaklines with the Helmholtz decomposition and spatially with K-means clustering by using Procruster distance.

Spatial clustering can also be applied directly on the optical flow to detect large groups and to study how they interact [11]. The orientation, position, size and displacement of groups are used to model force fields. In particular, the size is used to provide a degree of importance to the groups: the bigger the size, the more representative the group. This type of modeling enables the detection and prediction of interactions (e.g., merging, splitting) among groups relying on sudden variations in the direction of their movements. In order to improve the reliability of the method, the temporal consistency of group clusters can be employed by associating neighboring cluster centroids over time [3]. A temporally consistent motion can be used to detect typical motion patterns and dynamic motion patterns (i.e., position, relative distances among groups, velocities, orientations) can be classified and, similarly to [11], can be used to detect interactions such as merging, splitting or evacuation.

2.3 Object-based methods

Object-based methods for recognizing interactions use individual patterns or body parts. Individual patterns define spatial locations that can be computed using trackers. Body parts are used to infer details of people such as gaze or pose.

2.3.1 Individual patterns

Methods for detecting interactions based on individual motion patterns rely on automatic detection and tracking algorithms [2, 8, 14, 51, 65] or on manually annotated (ground truth) trajectories [43, 59, 63].

Automatically generated trajectories can be used to localize groups of people [9, 17, 51, 52]. On the one hand, groups can be treated as latent variables that are calculated from joint-trajectory information [52]. Interactions are modeled with a third-order Conditional Random Field (CRF), which considers reflexivity, symmetry and transitivity. These properties are learned using color and motion features. On the other hand, interactions can be modeled with the proxemic theory [8, 24]. Firstly, candidate interacting people are extracted by checking whether every pair of trajectories has 30 % of locations within the interaction boundaries defined by the proxemic theory [15]. Then, interactions are detected by analyzing similarities among proxemic distances and employing Needleman–Wunsch algorithm to obtain the final solution [48]. This algorithm is used because it is robust to fragmented trajectories. Interactions can also be studied at group level, where group interactions (e.g., dispersion, formation, following, meeting of groups) can be modeled using single models for each type of interaction, as a probabilistic grouping strategy applied to track pairs and solved with a graph-based approach [9]. The measurements for the group models are the distance between each pair of individuals, velocity and track history. The approach also employs a model for predicting people's locations in order to enable real-time processing of interactions. Furthermore, interactions can be modeled as sub-interactions to recognize complex activities. Sub-interactions are assumed to be continuous uninterrupted activities and their combination leads to a more flexible approach to model behavior changes and is more adaptable to different contexts [63, 65]. For example, it is possible to consider interactions between different subjects such as cars and people [68]. Sub-interactions (e.g., cars stopping, cars turning right, pedestrians crossing a street) are defined in order to detect abnormal interactions such as jay-walking. Since the motion caused by the same type of sub-interactions often co-occur in the same video, the features of moving objects are treated as words in hierarchical Bayesian models. When people interact, sequences of sub-interactions can be observed such as two individuals approaching a common location then meet and go separately, or approaching then meet and go together. These are example of interactions with minor variations in the temporal evolution of the action. In order to detect these sub-interactions, Coupled Hidden Markov Models (CHMM) are employed [49]. Alternatively, sub-interactions can be combined with a modified factorial HMM [63]. The method utilizes a network of dynamic probabilistic models (NDPM) for the representation of complex patterns with a combination of simple sub-interaction models. With NDPM, the extension to new complex interactions can be simply made by constructing new dynamic models and linking them to the main system. This favors the extension of the overall model in the case of newly modeled sub-interactions.

Annotated sequences are also used to demonstrate the validity of interaction models [43, 59, 60, 63]. The online inference of groups of people can be performed by using an iterative algorithm for error minimization based on SFM [59]. This method introduces a group force into the SFM in order to model human interactions within the same group. The algorithm predicts locations of people using the SFM with and without the group force, and calculates the location error with respect to real observations. People who provide a small error when the group force is taken into account are considered to be interacting. This algorithm fails when people not belonging to the same group are close to each other. Heuristics can improve the group recognition reliability [43] when people cross a group and people approach a group of stationary people to become part of it. Alternatively, groups can be detected using a Structural Support Vector Machine (S-SVM) classifier [60]. The learning framework exploits annotated data to deduce a sociologically motivated distance measure proposed by Hall's proxemics [15, 24] and Granger's causality [21]. Proxemics can also be used to define social constraints embedded in models operating in a feature space defined by the location and velocity of each person in each frame [69]. If groups are represented as components of a mixture model, and people are seen as observations for the mixture model, then a Dirichlet Process Mixture Model can be employed for group detection. Groups of interacting people can be associated to group activities, such as roles in sports [33]. Annotated trajectories are used to get target locations and to enable reliable extraction of Histogram of Oriented Gradients (HOG) [16] from the bounding boxes to infer individual social roles. Social roles are then distinguished into unary and pairwise. Unary roles define the social properties of each individual, whereas pairwise roles define the dependencies between pairs of individuals. Their combination is used to detect activities during a game: a defender defends against an attacker (man-marking), defenders defend space or against an attack play.

2.3.2 *Body parts*

Human interactions are also studied by looking at body-poses, such as head pose [14, 37] or gaze [19]. For example, people's trajectories along with head orientation are automatically computed to detect interactions by a statistical analysis of F-formations [14]. The concept of F-formation, as defined by Adam Kendon, states that: "*an F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access*" [27]. F-formations can be detected based on Hough voting [35] by weighting the extracted features (person location and head orientation) in order to provide locations on the image plane for candidate interactions. Head poses can also be exploited to understand group interactions in a conversation and can be inferred using two types of time-varying descriptors: per-agent descriptors that encode the appearance and/or motion of each individual, and relative pairwise descriptors that encode the appearance and/or motion of each individual jointly to another [37]. The method uses features from trajectories to locate people (e.g., position, velocity, acceleration, histogram of flow) and appearance descriptors (e.g., HOG [16], SIFT [41]) to model the interactions, and the recognition is performed using a combination of sub-interactions. From a first-person view (i.e., glasses with embedded camera) detection and tracking of faces can be used to infer the 3D gaze directions [19]. 3D gaze directions enable detection of the focus of attention (common location of gaze) of people and, along with the head locations and movements, to identify social roles within a group. The method is based on Markov Random Fields (MRF), and relies on the likelihood that a person is looking in a certain location of the 3D space given where the other faces are looking. Alternatively, the head pose information can be modeled by representing the visual field of a person inside a scene with a 3D polyhedron. By looking at the overlap of visual fields of candidate interacting people it is possible to infer whether social interactions are occurring [17].

2.4 Datasets

Datasets for the study of human interactions may contain crowds ranging from high-density to low-density. Chaquet et al. [10] presented a survey on datasets for people's actions and activity recognition, and Borges et al. [5] a survey on human behavior understanding. However, both papers list datasets that do not contain high-density (crowded) scenes. We present a list of datasets that range from mid- to high-density scenes (Table 1) and categorize them into outdoor, indoor and both.

Outdoor datasets include `Student003` that is [59] composed of one clip of 5,400 frames to study social interactions (group formations). Manual annotations of people's locations and groups are available. The `BIWI Walking Pedestrian` dataset is also used to study social interactions [52] and consists of two clips composed of 12,974 and 19,350 frames. Manual annotations of people's locations and groups are available for both datasets. `CoffeeBreak` serves for checking the ability in detecting social interactions (i.e., people chatting). The dataset is captured by two cameras and annotations of two sequences indicating groups present in the scenes are provided [14]. `PETS2009` is also used to detect panic situations as well as group interactions (e.g., merging or splitting) [62]. Manual annotations of people's locations and interactions are available. `Friends Meet` is used to study social interactions [2]. The dataset is composed of 28 synthetic clips of 200 frames each and 15 real clips of different

Table 1 Summary of available datasets for interaction analysis

	Name	Link	H-GT	I-GT	A
Outdoor	Student003	http://cmp.felk.cvut.cz/~sochmj1/	✓	✓	
	BIWI Walking Pedestrian	http://www.vision.ee.ethz.ch/datasets/index.en.html	✓	✓	
	CoffeeBreak	http://profs.sci.univr.it/~cristanm/datasets/CoffeeBreak/index.html	✓	✓	✓
	PETS2009	http://www.cvg.rdg.ac.uk/PETS2009/a.html	✓	✓	
	Friends Meet	http://www.iit.it/en/datasets/fmdataset.html	✓	✓	
	MPR	http://www.mockprisonriot.org/		✓	
	BEHAVE	http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/	✓	✓	
	UT	http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html		✓	
	VIRAT	http://www.viratdata.org/		✓	
	DARPA Mind's Eye	http://www.visint.org/datasets		✓	
	ISSIA	http://www.ino.it/home/spagnolo/Dataset.html	✓		
Indoor	Train Station	http://www.ee.cuhk.edu.hk/~xgwang/grandcentral.html			✓
	JAIST	http://www.jaist.ac.jp/~chen-fan/multivision/jaistmvsdb.html	✓	✓	✓
	Harmes	http://goo.gl/yu82B5			✓
	APIDIS	http://www.apidis.org/Dataset/	✓		✓
Both	Collective Motion Database	http://mmlab.ie.cuhk.edu.hk/archive/project/collectiveness/dataset.htm			✓
	UMN	http://mha.cs.umn.edu/Movies/		✓	

Websites last accessed: April 2014

H-GT ground-truth locations of people, *I-GT* ground-truth of interactions, *A* automatically precomputed trajectories or detections

lengths (between 30 and 90 s). Manual annotations of people's locations and interactions are available. MPR is provided by the Mock Prison Riot and contains group interactions (gang activities), such as agitated arguments, fights, contraband exchanges [9]. BEHAVE is used to detect interactions (e.g., approach, meet, split, chase, fight, etc.) among several people (e.g., around four or five) and consists of four clips from two different views. Manual annotations of people's locations and interactions are provided. UT interaction in used to detect interactions (e.g., hand shaking, hugging, kicking, etc.) between two or four people [55] and consists of 20 clips with different backgrounds, scales and illuminations. Manual annotations exists for the type and location of the interactions. CASIA is used for the detection of interactions between two individuals (e.g., rob, fight, follow, follow and gather, etc.) [26]. VIRAT is an outdoor dataset recorded from both ground and aerial cameras and contains interactions between human–object (e.g., a person closing a trunk or entering in a vehicle) and human–human (e.g., two people loading a trunk) [47]. Manual annotations of the location and type of interactions are provided. DARPA Mind's Eye contains both human–object and human–human interactions. ISSIA is a soccer dataset recorded from multiple cameras [66]. Annotations of people's locations are provided for a 2-min clip.

Indoor datasets include The Train Station that is used to study collective human behaviors [71] and consists of a clip composed of 50,010 frames along with precomputed trajectories. JAIST is used to study group interactions indoors [12]. Manual annotations of people's locations and interactions are available. Harmes contains pedestrians walking through a corridor with a bottleneck at the end of the corridor. The scene is recorded in laboratory conditions using two overhead cameras with a small overlapping field of view [31]. APIDIS is a basketball sport dataset recorded from multiple cameras [12]. Manual annotations and detections of people's locations are provided.

Both outdoor and indoor datasets include The Collective Motion Database [72] that consists of 413 video clips from 62 crowded scenes and each clip has around 100 frames along with precomputed trajectories. Clips to study human behaviors are included in Gettyimages, BBC Motion Gallery, Thought Equity Motion or Youtube as done in [61]). The UMN dataset is used to study interactions in order to detect panic situations [45]. It is a single clip composed of sub-clips and the annotations are related to the events of panic situations.

3 The role of rendezvous areas for the prediction of human interactions

3.1 Social reasoning

Features measuring interpersonal distances (proxemics) and their relative variations over time are key for the study of social interactions. When people share a space and maintain it over time, their reciprocal distance and body orientation can be used to understand when they are interacting. These spatio-orientational arrangements (formations) require the cooperation of all participants. Proxemic theory [24] and formation models (e.g., F-formation) [28] are in fact used in computer vision to infer people's behaviors by means of spatio-temporal features. Both theories consider spatial layers around each individual to define spaces of interactions each of which has a sociological meaning. When people interact they need to traverse these layers. For example, in an F-formation if a person wants to join an ongoing group, they have to be in the p-space (i.e., area where the participant's bodies and also personal things such as handbags are placed) and seek to join the group, then the other participants have to alter their spacing in order to let the individual be part of the group.

In a dynamic environment, where there are no current formations, future participants will generate formations to interact (e.g., communicate). In a simple structured scenario with no obstacles, the formation will take place in an area chosen by the participants such that the effort spent to get there is minimum (shortest path). We exploit this idea in real scenarios in order to find candidate areas of group formations (rendezvous areas) before the formations occur.

3.2 Rendezvous areas for tracking and social analysis

Areas where people consistently tend to converge are likely to be locations of interactions (e.g., group formations). We will show how areas of interactions (rendezvous areas) enable inference of structures in both visible and non-visible areas (i.e., outside the field of view of a camera). The approach we present in Sect. 4 is a simple but effective idea that allows one to automatically infer dynamic (temporary) goals or static (highly likely) areas where, for example, failures can occur.

Different applications with time-critical constraints can benefit from a method capable of detecting rendezvous areas. Liu et al. [40] exploit focus areas to improve people tracking. In these areas the density of people is expected to be high, and it is more likely that tracking failures occur. Their tracking approach is based on a graph network and trajectories are inferred with the optimal solution that associates all the detections in a video sequence. The focus areas are used to predict people's motion in order to activate procedures for avoiding failures. Motion prediction methods often use manually labeled goals to meet the assumption that people have goals driving their motion and future movements (~ 0.4 s) can then be inferred as a function of these goals [51]. Also, prediction methods generally assume that moving people do not abruptly change their velocity or locations [4]. This can be valid for short-term predictions, but it is not always valid for longer ones (>1 s), especially in panic situations or mid-density crowds. A long-term prediction can enable the actuation of preemptive systems for the reinforcement of target features, for example, by strengthening the color feature model before the occurrence of an occlusion in order to perform track re-association in the case of a failure. However, to the best of our knowledge, explicit long-term predictive models have not yet been extensively employed for the unsupervised inference of areas of likely tracking failures.

Crowd simulation algorithms aim to achieve real-time performance by employing prediction of people's locations in order to infer likely collisions [70]. Simulators usually provide a set of candidate people states that represent their locations and velocities. The states that are more likely not to cause collisions ahead of time are selected by the algorithm. Moreover, socially aware robots have to be capable of understanding social behaviors in order to integrate themselves into human-populated environments [30]. For example, Papadakis et al. [50] introduced a social mapping that models context-dependent human spatial interactions. Authors represent the social space as a probability density map and, using isocontours, they define zones where people interact (social zones). The navigation of robots in socially aware environments can embed predictive models of interactions to improve online path planning methods in social spaces. Most of the methods surveyed in Sect. 2 are formulated as offline or buffered [8, 15], and are not suitable for time-critical applications, whereas those online are still in an early stage [43]. The computation of rendezvous areas can allow one to infer people's "intentions" for group formations and can be performed with a delay shorter than the time to the event. The intention can then be validated using the observed tracks. Moreover, an interaction localization approach composed of prediction and update stages would allow one to achieve robustness to noise. Similarly to the linear evolution model of a particle filter, the more generic the prediction model, the more applicable it is to different scenarios.

4 Prediction of human interactions

We exploit motion orientation to predict rendezvous areas, i.e., the areas where people are expected to meet or to approach a specific object such as a door, a bench or a kiosk.

4.1 Prediction of rendezvous areas

Let $V = v^{1,T}$ be a video sequence of length T frames and $P^t = \{p_n^t\}_{n=1}^{N^t}$ be the set of N^t people's locations at time t,

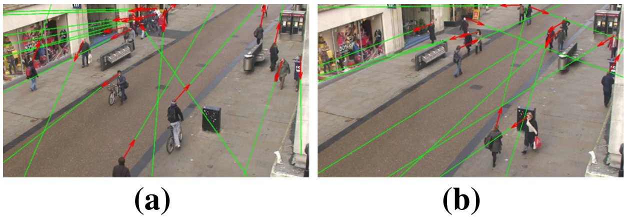

(a) (b)

Fig. 3 Sample frames representing the prediction lines (*green lines*) drawn from displacement vectors. The *red arrows* intuitively represent the direction of motion

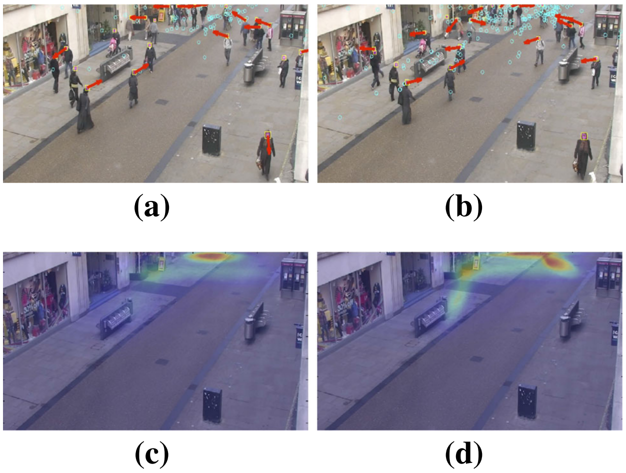
where p_n^t is the nth individual. Let $\phi_i^t = (x_i^t, y_i^t, u_i^t, v_i^t)$ be the generic ith displacement vector between two consecutive frames, v^{t-1} and v^t, defined as a four-vector component where (x_i^t, y_i^t) is the position and (u_i^t, v_i^t) represents the components along horizontal and vertical axes, respectively. We denote the set of displacement vectors as $\Phi^t = \{\phi_i^t\}_{i=1}^{I^t}$, where I^t is the total number of displacement vectors at time t. Ideally $N^t = I^t$, but problems of occlusions and mis-detected or false-positive targets may lead to $N^t \neq I^t$. Displacement vectors can be calculated through optical flow [64] or with multi-target tracking algorithms [1].

Let us consider the people's motion direction as the extension of the displacement vector. Given the set Φ^t, we extend the direction using prediction lines [53]. We define a prediction line that lies on each ϕ_i^t, such that

$$\begin{cases} \beta_{i,1}^t & = & \tilde{x}_i^t \\ \gamma_{i,1}^t & = & \tilde{y}_i^t \\ \beta_{i,2}^t & = & \tilde{x}_i^t + \tilde{u}_i^t \\ \gamma_{i,2}^t & = & \tilde{y}_i^t + \tilde{v}_i^t, \end{cases} \tag{1}$$

where $\beta_{i,1}^t, \beta_{i,2}^t, \gamma_{i,1}^t, \gamma_{i,2}^t$ are the parameters that define the straight line passing through two points. For each constructed prediction line we consider the segment that belongs to the position of the displacement vectors onward with respect to its orientation. Figure 3 shows representations of displacement vector extensions.

A rendezvous area is at the convergence of multiple extended directions of people's motion. Prediction lines are utilized for the localization of rendezvous areas. Let us define $\Psi^t = \{\psi_j^t\}_{j=1}^{J^t}$ as the set of J^t intersections at time t, where $\psi_j^t = (\hat{x}_j^t, \hat{y}_j^t)$ is the jth intersection of two prediction lines and $\hat{x}_j^t, \hat{y}_j^t$ are the coordinates on the horizontal and vertical axes, respectively. We collect the intersections ψ_j^t for a certain interval of time τ, in order to consider a representative and meaningful set of intersection points. Hence, let us define the buffer $B^{t-\tau,t} = \{\Psi^{t'}\}_{t'=t}^{t-\tau}$ as the collection of intersection points within the temporal window τ.


(a) (b)

(c) (d)

Fig. 4 Rendezvous points on the image plane generated: **a** by intersections of prediction lines and **b** accumulated over time. *Red arrows* indicate the principal direction of people's motion. *Cyan circles* represent the accumulated intersections over time. **c**, **d** Evolution of the GMMs fitted on the respective intersection points

When a large group of people moves in the scene, intersection points may appear in different locations of the reference plane (image plane or ground plane) (Fig. 4). Instantaneous people's motion analysis (i.e., between two consecutive instants) is noisy due to the non-rigidity of the body, and this affects the estimation of the principal direction of motion. Such noise is mainly generated by the *bouncing* movement of people while they walk. This problem may compromise the estimation of the rendezvous areas by generating intersection points in incorrect locations. For example, in Fig. 4 we can identify areas with sparse and high concentrations of intersection points over the image.

In order to obtain a smooth and continuous spatial distribution of the rendezvous areas and to enhance areas with high concentration of intersection points, we fit Gaussian Mixture Models (GMM) on such intersection points. GMMs enable us to deal with reliable estimations of spatial concentrations of the intersection points while remaining robust to the noise caused by spurious intersections. GMMs are fitted on the accumulated intersection points $B^{t-\tau,t}$ and represent the rendezvous areas for instant t. In particular, for each t we use an unsupervised GMM method [6] to describe the spatial distribution of the accumulated points. A GMM is defined as a weighted sum of Gaussian densities given by

$$f(B^{t-\tau,t}) = \sum_{m=1}^{M^t} \pi_m^t \mathcal{N}\left(B^{t-\tau,t} | \sigma_m^t, \mu_m^t\right), \tag{2}$$

where M^t is the order of the mixture at time t, which can be estimated with the Minimum Description Length (MDL) criterion [6]. μ_m and σ_m denote the mean and covariance parameters, respectively. π_m is the weight of the mth fitted Gaussian. The parameters and weights of the mixture are estimated with the Expectation–Maximization (E-M) algorithm [6].

For each t, due to the noisy principal direction, the quantity and location of intersection points $B^{t-\tau,t}$ varies substantially (Fig. 4a, b), leading to variations of the Gaussian mixture distributions (Fig. 4c, d). To increase the robustness to such noise, we perform temporal filtering of the modeled distributions $f(B^{t-\tau,t})$. The distribution $\mathcal{F}^t$ represents the final distribution of the dynamic rendezvous points at time t and is given by the temporal mean filtering of $f(B^{t-\tau,t})$ within the temporal window τ: $\mathcal{F}^t = g(f(B^{t-\tau,t}))$, where $g(\cdot)$ represents the operation of temporal mean over the window τ.

The distribution $\mathcal{G}$ that represents the static rendezvous points is given by the temporal accumulation of the dynamic centers of attention:

$$\mathcal{G} = \sum_{t=1}^{T} \mathcal{F}^t. \tag{3}$$

4.2 Discussion

Limitations of the method proposed in this paper include sensitivity to motion variations caused by the non-rigidity of people's movements and the fixed length of the prediction lines. The latter can be improved by modulating the length of prediction lines as a function of people's velocity (e.g., first order linear model). Another element of noise is the perspective distortion introduced by the cameras. Moreover, when a video sequence is short and the motion pattern of each individual is constrained by other people (e.g., a large crowd moving in one direction), the rendezvous area converges to an entry/exit zone. A possible way to model such a scenario is by considering social forces [25]. This will modify the orientation of the prediction lines as a function of the current and future people's locations. For example, each point along the prediction line can be a function of other points on the same prediction time. Therefore, each prediction line will become a curved line accounting for neighboring lines. The function will be constructed following the social force model, which is largely employed in the literature to model people's behavior in crowds. A similar idea was presented by Pellegrini et al. [51].

Unlike Poiesi et al. [53], the method proposed in this paper uses a more generic modeling that allows its real-time employment and it is suitable for multimodal distributions (i.e., multiple areas of interest). In our 2010 work we aimed to estimate the location of the ball in basketball matches by analyzing the collective movement of the players. We used a path-selection algorithm and a Kalman filter to find the most likely location over time and to filter out noisy measurements. Such a procedure requires the collection of all candidate ball locations throughout the sequence (offline) and it is not suitable for multimodal distributions (i.e., only one area of interest can be considered).

5 Results

5.1 Experimental setup

We validate the proposed method on six surveillance scenarios. In four of them people's locations are on the image plane, whereas two use the ground-plane locations.

The first dataset is from Camera 1 of the London Gatwick airport (London, UK) dataset.[2] The video is composed of 2,400 frames of size 720 × 576 pixels at 25Hz. We refer to this dataset as Trecvid-Cam1. The second dataset is iLids-Medium[3] and it is composed of 4,800 frames of size 720 × 576 pixels at 25Hz. The third dataset is TownCentre[4] and it is composed of 4,500 frames of size 1,920 × 1,080 pixels at 25Hz. People's locations are manually annotated at the head position. The fourth dataset is Students003 and it is composed of 5,400 frames of size 720 × 576 pixels at 25Hz. People's locations for this dataset are manually annotated and we use the projection on the ground plane. The fifth dataset is PETS2009,[5] which is composed of three videos, namely S2L1, S2L2, S2L3, of frames 794, 435, and 239, respectively, and of size 768 × 576 pixels at 7 Hz. People's locations for this dataset are manually annotated [1] and, as for Students003, we use the projection on the ground plane. The principal direction is calculated using the location difference of the annotations frame-by-frame. The sixth dataset is Train Station of New York Grand Central recorded from a high top-view of size 720 × 480 pixels at 23 Hz and we use the first 5,000 frames. We refer to this dataset as TrainStation and people's locations are extracted using the KLT tracker provided by [71]. The displacement vectors are calculated by a frame-by-frame difference of people's locations and $\tau = 15$.

[2] iLIDS, Home Office multiple camera tracking scenario definition (UK), 2008.

[3] http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html. Last accessed: December 2013.

[4] http://www.robots.ox.ac.uk/ActiveVision/Research/Projects/2009bbenfold_headpose/project.html. Last accessed: December 2013.

[5] http://www.cvg.rdg.ac.uk/PETS2009/a.html. Last accessed: December 2013.

We qualitatively and quantitatively evaluate the results. The former evaluation is performed via analysis of visual sample results, the latter by quantifying similarities as follows. We firstly evaluate the method with the aim of understanding its prediction capability. Similarly to [29], we calculate the log-likelihood function between people's positions in multiple instants ahead of time, i.e., $P^{t'}$ with $t' = t + \gamma$ where $\gamma \geq 0$, and the predicted rendezvous areas at time t, i.e., $f(\Omega^t)$. The log-likelihood function is calculated as

$$\ln p(P^{t'}|\pi^t, \mu^t, \sigma^t) = \sum_{n=1}^{N^{t'}} \ln\left(\sum_{m=1}^{M} \pi_m^t \mathcal{N}(P^{t'}|\sigma_m^t, \mu_m^t)\right), \qquad (4)$$

where π^t, μ^t, σ^t are the parameters estimated with the E-M algorithm from Eq. 2.

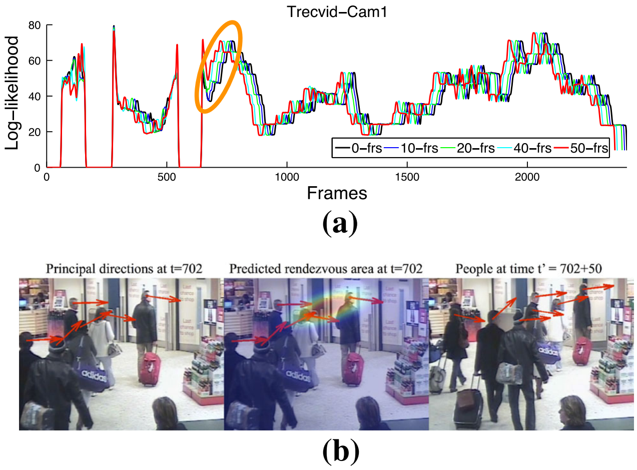
Next we assess the robustness of the method by adding noise to people's locations and comparing whether two estimated rendezvous area distributions are similar to each other. By doing this we can understand the extent of bias of the noise on the results. We use the Symmetric Kullback–Liebler divergence, SKL [57], which is a common way to measure the distance between two distributions. Let $\mathcal{G}$ and $\mathcal{G}'$ be the two distributions to compare. The SKL divergence is calculated as follows:

$$\mathrm{SKL}(\mathcal{G}, \mathcal{G}') = \left|\frac{1}{\mathrm{K}}\sum_{\mathrm{k}=1}^{\mathrm{K}} \ln \mathcal{G}(\mathrm{k}) - \frac{1}{\mathrm{K}}\sum_{\mathrm{k}=1}^{\mathrm{K}} \ln \mathcal{G}'(\mathrm{k})\right|, \qquad (5)$$

where K is the number of sample data generated from $\mathcal{G}$ and $\mathcal{G}'$.

5.2 Dynamic rendezvous areas on the image plane

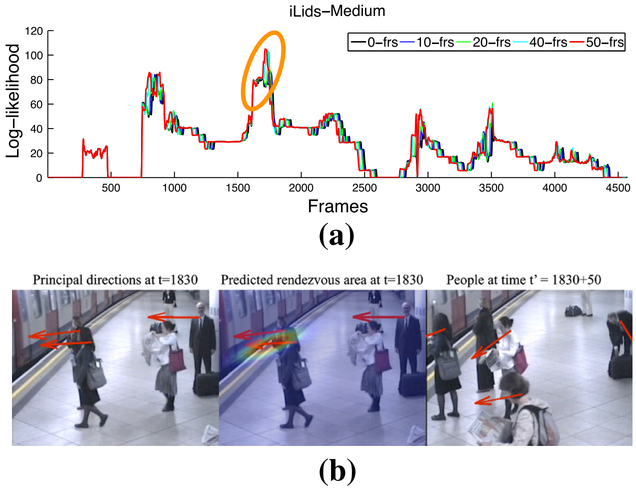
Dynamic rendezvous areas on the image plane are evaluated on Trecvid-Cam1, iLids-Medium and TownCentre. Figures 5, 6 and 7 show the results of Trecvid-Cam1, iLids-Medium and TownCentre, respectively. Since the camera field of views are different between the datasets, we variate the value of γ. The graphs for Trecvid-Cam1 and iLids-Medium are generated for same values of $\gamma = 0, 10, 20, 40, 50$. In TownCentre the field of view is wider and we use $\gamma = 0, 40, 50, 60, 80$. The graphs show the trend of the likelihood functions over time. The value of the likelihood is large when the spatial distribution of people matches correctly with that estimated with GMMs. The trend of likelihood functions at different time delays (according to γ) shows that the estimation of the rendezvous areas can enable the prediction of the distribution of people ahead of time. From the graphs, we analyze the results when the trends of the likelihood function show consistency over time.



(a)

(b)

Fig. 5 Log-likelihood functions of the predicted rendezvous area calculated on Trecvid-Cam1 dataset over time. Each function represents the likelihood between the rendezvous area $f(\Omega^t)$ and the spatial distribution of people $P^{t'}$ ahead of time



(a)

(b)

Fig. 6 Log-likelihood functions of the predicted rendezvous area calculated on iLids-Medium dataset over time. Each function represents the likelihood between the rendezvous area $f(\Omega^t)$ and the spatial distribution of people $P^{t'}$ ahead of time

As far as Trecvid-Cam1 dataset is concerned, Fig. 5 shows a group of people moving towards the door at the end of the corridor. The estimation of the rendezvous area provides a higher likelihood when the distribution of people is considered 50 frames ahead. In fact, the example shows that the concentration of people later in time is higher in the predicted area. The predicted rendezvous area in Fig. 5 is located slightly on the left with respect to the door which the group of people is heading toward. This fact is mainly due to the perspective of the camera, which biases the perception of the people's principal direction.

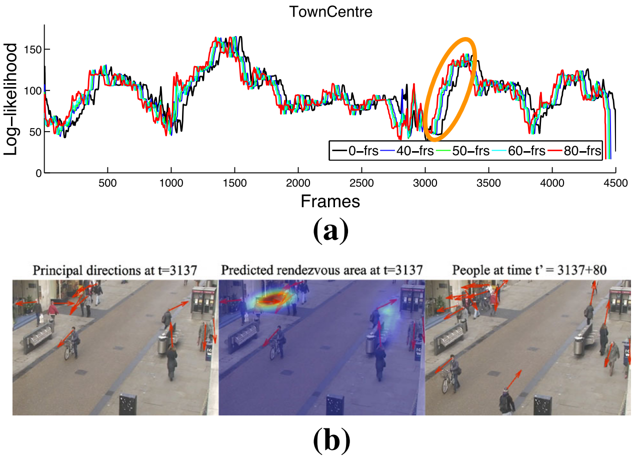


Fig. 7 Log-likelihood functions of the predicted rendezvous area calculated on TownCentre dataset over time. Each function represents the likelihood between the rendezvous area $f(\Omega^t)$ and the spatial distribution of people $P^{t'}$ ahead of time

The rendezvous area on iLids-Medium (Fig. 6) is predicted by the interest of people when the train is approaching the station. The attention of people is in fact focused on this element of interest and the likelihood function has a higher value for $\gamma = 50$. The third case is analyzed on TownCentre. Figure 7 shows a group of people on the top-left corner that is converging in the area highlighted on the pavement. The predicted location is well localized and it is also clear from the graph how the likelihood function has a higher value for $\gamma = 80$. The predicted rendezvous area on the right-hand side of the image is due to the motion generated by the two people walking one behind the other.

5.2.1 Dynamic rendezvous areas on ground plane

We perform a qualitative evaluation on Students003[6] (Fig. 8) and PETS2009[7] (Figs. 9 and 10). Some sample frames are modified by including a border around the image in order to analyze rendezvous areas outside the field of view.

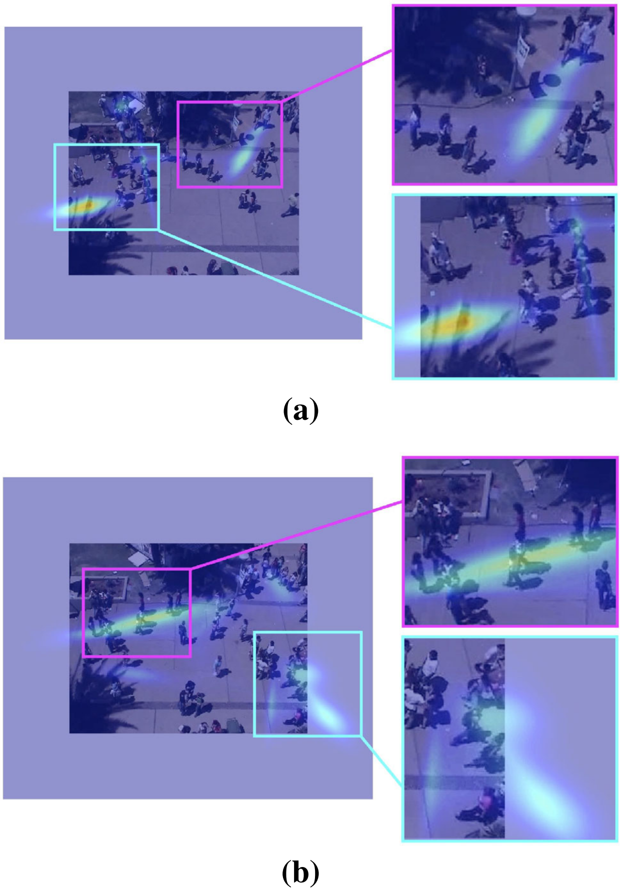


Fig. 8 Sample results of predicted rendezvous areas on Student003. **a** Two rendezvous areas predicted within the field of view, whereas **b** one rendezvous area is predicted outside the field of view

Figure 8 shows two examples of rendezvous areas generated by groups of people converging on the same locations. Within the magenta quadrant of Fig. 8a the convergence location is modeled with a mixture of Gaussians; there is a group of people moving from the top to the bottom, a group moving from the bottom-left to the top-right and a person from the right to the left. Their movement leads to motion patterns that create a set of intersection points that in turn generates the mode of the distribution. A similar situation is occurring on the left-hand side of the image (cyan quadrant). Groups of people are converging in a common area, but since the number of people involved is greater than the top-right one, the rendezvous area on the left generates a mode of the distribution with a higher amplitude. The shape of the modes depends on the spatial distribution of the intersection points, elongated shapes are generated by the intersection points distributed along prediction lines (one component of the covariance matrix of GMMs has a bigger magnitude than the other). These are likely to be generated by people converging from parallel but opposite directions, whereas circular modes are more likely to be generated by people converging from orthogonal directions (components of the covariance matrix have similar magnitude values). In fact Fig. 8b contains an elongated mode (magenta quadrant) that is generated by converging people with almost "parallel" directions. Figure 8b also shows a rendezvous area outside the field of view (cyan quadrant) that is generated by exiting people and gives meaningful information about the structure of the

[6] Video results on the full sequence can be found here: ftp://motinas.elec.qmul.ac.uk/pub/ra_results/students003_border.zip.

[7] Video results on the full sequence can be found here: ftp://motinas.elec.qmul.ac.uk/pub/ra_results/pets2009_noborder.zip.

environment. By looking at the video of the results we can notice that people tend to arrive from the walk path on top of the scene and exiting at the bottom-left corner. At the same time people coming from the right-hand side of the scene exit at the same corner. This generates rendezvous areas outside the field of view which enable us to guess a likely structure of the environment (like a possible unobservable walk path) or to infer the intention of the converging people to go in the same direction.

Figures 9 and 10 show examples of the dynamic rendezvous areas extracted from PETS2009, in particular from S2L1 (Fig. 9) and S2L2 (Fig. 10) sequences. In Fig. 9 there is a situation where people have just met and the rendezvous area is clearly visible in their location. The area remains visible also when the interaction has occurred, since the accumulation buffer still contains the intersection points and takes almost its duration (τ) to become empty. No other rendezvous areas are detected in the image, even though there are moving people. The difference between Fig. 9a and b is that the former is calculated using ground-plane people's locations and the intersection points are extracted on the ground plane. In the latter, the rendezvous area is calculated using image-plane people's locations. These examples show how the ground-plane locations allow more accurate estimation of rendezvous areas. In Fig. 9b the area is larger because the straight lines of the people coming from the lower part of the street intersect that of the person not involved in the meeting (right-hand side in the cyan quadrant). This is due to the perspective distortion, whereas in real-world coordinates (ground plane) their directions do not intersect. In Fig. 10a rendezvous areas are calculated using ground-plane people's locations and two rendezvous areas are detected. In one case (cyan) people on the right-hand side are intersecting another person coming from the grass field before exiting the view-point. In the other case (magenta), people coming from the top of the scene are intersecting near the street lamp. The difference in intensities between a rendezvous area with a meeting event that actually occurred (Fig. 9a) and a rendezvous area where a meeting did not occur are interesting to observe (Fig. 10a, b—cyan rectangles). The intensity of the former is much higher than that of the latter. In fact, in the latter case people do not meet because those coming from the right-hand side changed direction and went backward.

Later in this section, we will show that by accumulating dynamic rendezvous areas over time we can infer additional information about the structure of the environment.

5.2.2 *Static rendezvous areas on image planes*

Static rendezvous areas, calculated with Eq. 3, are qualitatively evaluated by comparing the results with a method similar to [44]: GMMs are applied to cluster people's trajectories in order to infer areas of interest (entry/exit points). Differently from [44], we use ground-truth trajectories instead of automatically computed ones. In principle, the higher the concentration of people's locations in a particular area, the higher the people's interest in such an area should be. Figure 11a–c shows the accumulated people's locations and Fig. 11d–f shows the respective fitted distributions.

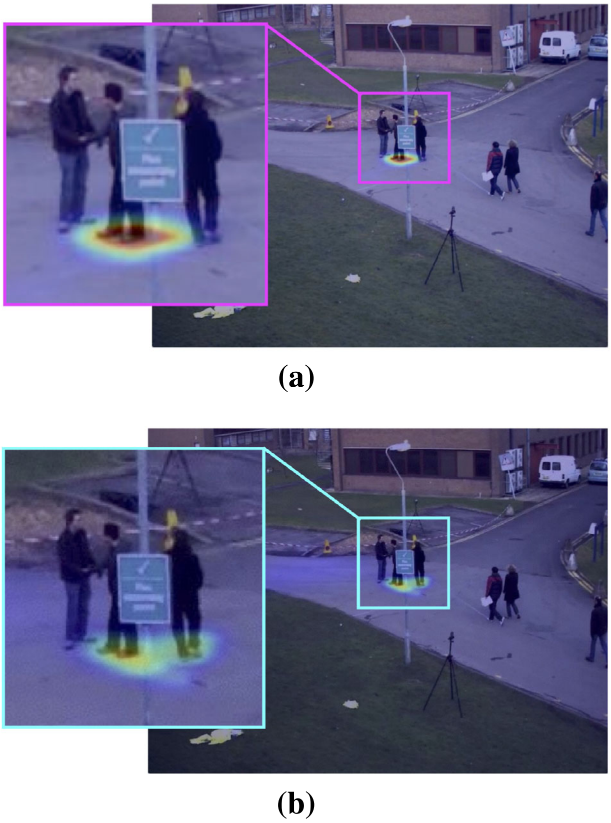


Fig. 9 Sample results of predicted rendezvous areas on PETS2009 (S2L1) in the case of "people meet". The example shows the difference in predicting a rendezvous area when **a** ground-plane locations or **b** image plane locations are used

In Trecvid-Cam1 (Fig. 11), three main areas can be identified: information desk (bottom-right with highest importance), kiosk (top-left with second most importance) and doors (top-center with lower importance). As far as the detected areas close to the information desk and to the kiosk are concerned, we believe they are not of high interest since there are no trajectories converging on those locations. However, the GMM modeling provides two modes in those locations given the continuous presence of operators. The exit point detected on the doors is the only correct one (Fig. 12a) and this is the only area detected by the proposed method.

In iLids-Medium only one area is detected using true people's locations (Fig. 11e). People from the scene pass

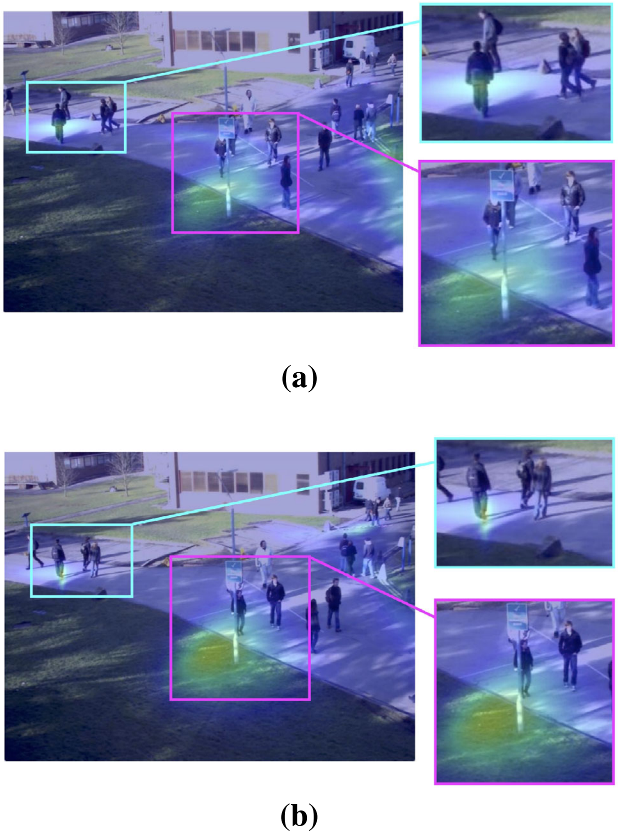


Fig. 10 Sample results of predicted rendezvous areas on PETS2009 (S2L2) using ground-plane locations. Two frames (**a** 171 and **b** 186) showing cases of rendezvous areas where people change direction prior to reaching the expected rendezvous area

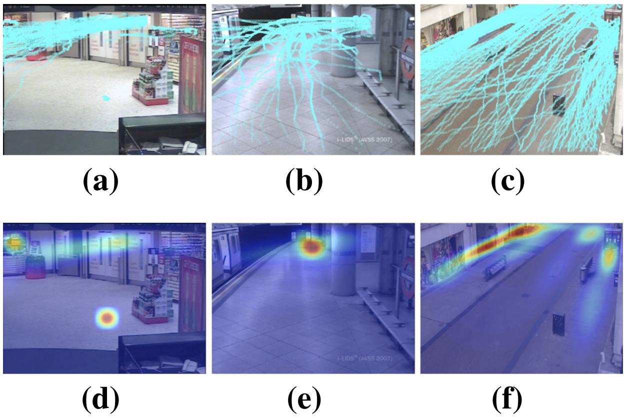


Fig. 11 Accumulated manually annotated people's locations over time for **a** Trecvid-Cam1, **b** iLids-Medium and **c** TownCentre datasets. High concentration areas obtained by fitting GMMs on the accumulated people's locations: **d** information desk, kiosk and exiting door; **e** corridor intersection; **f** sidewalks and upper entry/exit area

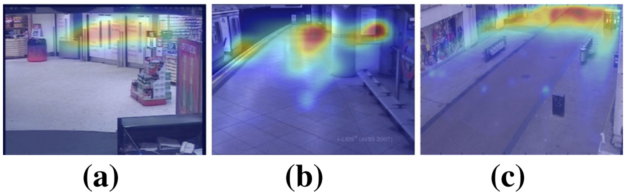


Fig. 12 Static rendezvous points on **a** Trecvid-Cam1 (exit door), **b** iLids-Medium (corridor intersection, train door, entry/exit corridor) and **c** TownCentre (upper entry/exit area) obtained by accumulating the dynamic rendezvous areas over time

with high frequency in this area and we can consider it of interest. Here, people coming from the right (entry/exit point of the train platform) cross people coming from either the bottom of the image and people alighting the train. Differently, from the results of the proposed method in Fig. 12b, three rendezvous areas are detected. In this case, the areas are more meaningful. The door of the train is an important rendezvous area. When people get on the train they all meet in the same area (at least those who have decided to enter from that door). The same area detected with the ground truth is also confirmed with the proposed approach. Finally, the exit point on the top-right of the image is correctly detected as a rendezvous area. People who are exiting and coming from a different location on the platform have motion converging in this area.

In TownCentre the accumulated people's locations are densely distributed along the left-hand pavement and this generates the modes of the distribution with the highest peaks (Fig. 11c). The results obtained with the proposed method (Fig. 12c) show the main rendezvous area at the top of the scene where people converge for exiting from the scene, whereas at the bottom no areas are detected since the trajectories are more sparse. Here, the perspective of the camera is a bias for both estimations. The trajectories appear to be highly dense in the top-left and the detected rendezvous areas are at the vanishing point of the scene.

5.2.3 Static rendezvous areas on ground planes

Figure 13 shows the comparison between areas extracted with people's locations (Fig. 13a, b) and estimated with the prediction method (Fig. 13c). We also remove the rendezvous areas within the field of view in order to highlight those outside. Static rendezvous areas are tested using Students003 and PETS2009 since it is possible to perform the analysis on the ground plane. Moreover, we quantitatively assess the robustness of the static areas in Students003 by introducing noise to people's locations.

From the results we can see that the main rendezvous area is located in the center of the scene (Fig. 13c). Looking at the modeled distribution of the trajectories in Fig. 13b this appears to be a reasonable result. Here, the main flows of people generate a distribution with a *quasi-*

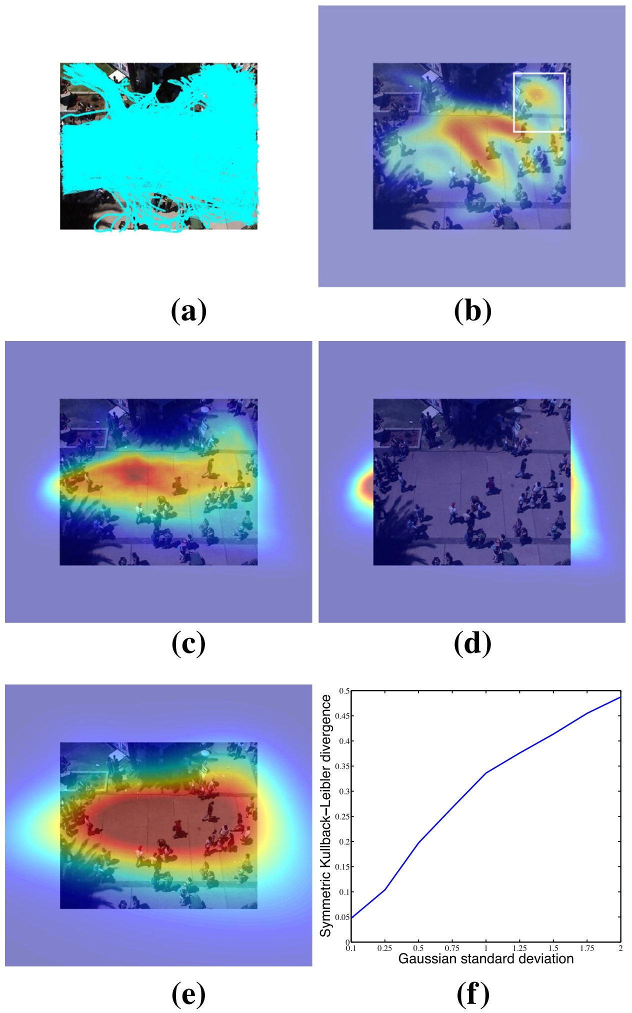


Fig. 13 Sample frames from Student003 with a border added to the frame to show the **a** density of manually annotated people's locations; **b** fitted GMMs on the people's locations; **c** static rendezvous points; **d** rendezvous areas outside the field of view; **e** static rendezvous area calculated by adding Gaussian noise (mean = 0 m and standard deviation = 2 m) to people's locations and **f** Kullback–Leibler divergence between **c** static rendezvous area without noise and static rendezvous areas with Gaussian noise with increasing standard deviation

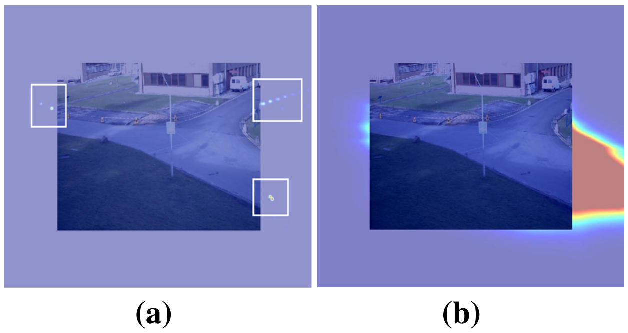


Fig. 14 Sample frames from PETS2009 with a border added to the frame to show static rendezvous areas outside the field of view in the case of the sequences (**a**) S2L1 and (**b**) S2L2

cross shape and the center coincides with the peak of the detected rendezvous area. In the same way as for the cases discussed in Sect. 6.2.2, the mode of the distribution in the top-right corner of Fig. 13b (white quadrant) is due to the accumulation of static trajectory points in that area. An interesting aspect to analyze is the minimum of the distribution located just below the mode (bottom of the white quadrant), whereas in the same location but in Fig. 13c there is activity in the rendezvous areas. We can infer that it is likely that the first intention of people coming from the main square and going toward the top-right is to stay in the middle of the path. However, due to the presence of the crowd, people have to deviate from the desired direction in order to avoid each other. Moreover, the extent of the minimum of the distribution can also be used to infer the distance that people maintain in order not to collide. Static rendezvous areas within the field of view in the case of PETS2009 are not as meaningful as in Students003 since the sequences are short and the motion patterns of the people across them are highly structured; for example, there can be sequences showing only a large group of people going from one side of the scene to another, or sequences with people moving randomly. We also use Students003 and PETS2009 to detect rendezvous areas outside the field of view (Figs. 13 and 14).

Rendezvous areas outside the field of view can be used to predict structures of unobservable areas or to detect exit locations. Figure 13d clearly shows how the two main exit locations are detected and also that on the right-hand side of the image the lower area is of more attention with respect to the higher one. By looking at the video sequence it is in fact possible to observe that the most common behavior involves people coming from the top and from the left part of the scene and converging to the bottom-right. From this we can infer that there should be a path in the bottom-right direction and, for this scene, it is of more interest to the people. Additionally, by looking at the modeled distribution it is also possible to infer an additional structure of the unobserved area as the bi-modality can indicate the presence of another path that is the straight continuation of the visible one. Conversely, there is a single rendezvous area outside the left-hand border since the path seems to be constrained by the tree at the bottom and the flower box at the top. Figure 14 shows estimations extracted from PETS2009 (S2L1 and S2L2). Figure 14a depicts the rendezvous areas and clearly shows the effectiveness of the method in detecting the areas outside the field of view. These areas are not as large as those of Students003 since the sequence is shorter and hence there is a lower density of intersection points. Conversely, in

Fig. 14b the intensity is higher, but the sequence used for the estimation is shorter (S2L2) and contains a large group of people exiting from the right-hand side of the scene and a few people exiting on the left-hand side.

The quantitative evaluation of static rendezvous area estimations, $\mathcal{G}$, is performed using the Symmetric Kullback–Leibler (SKL) divergence (Eq. 5). Starting from a $\mathcal{G}$ that is calculated without noise, we add to people's locations gradually increasing Gaussian noise (used to estimate the displacement vectors) in order to mimic a real scenario with automatically generated detections. The noise has zero mean, and the standard deviation is the same for x and y components and equal to 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2. The noise is applied on ground-plane people's locations and the values are in meter units. We measure the extent of uncertainty by calculating the SKL divergence between $\mathcal{G}$ computed without noise and each $\mathcal{G}'$ computed with noisy locations. The graph in Fig. 13f shows how the SKL divergence increases as the standard deviation increases. This is due to the noisy people's locations that bias the computed direction of people while introducing uncertainty in the estimation of the rendezvous areas. As a matter of the fact, from Fig. 13e we can see that the rendezvous area in the case of the highest standard deviation (i.e., 2) is larger than the case without noise (Fig. 13c). Therefore, the use of real detections can either affect the accuracy with which the intersection points (candidate convergence points) are accumulated in an area and the extent of the rendezvous area modeled with GMMs.

5.2.4 Rendezvous areas from top-down views with estimated motion patterns

TrainStation is used to evaluate the proposed method as a fully automatic pipeline.[8] The motion patterns are extracted with the KLT tracker provided by Zhou et al. [71].

Figure 15 shows two sample frames extracted from the video results where rendezvous areas are located in the middle of large converging groups. In this case, the largest convergence areas are located in the right-hand part of the scene (magenta quadrants) since the majority of the crowd is on that side. In Fig. 16b there is a large group of people converging to an exit location and it is clearly highlighted by the mode of the distribution. The white quadrants show the distributions of the rendezvous areas with clearly different shapes: in the case of the elongated shape we can observe that the people are vertically distributed and converging to a single direction, whereas in the case of the circular distribution people are more spread and converging from multiple directions.

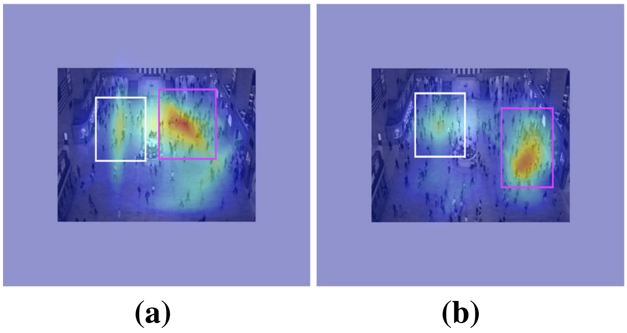


Fig. 15 Predicted dynamic rendezvous areas on TrainStation

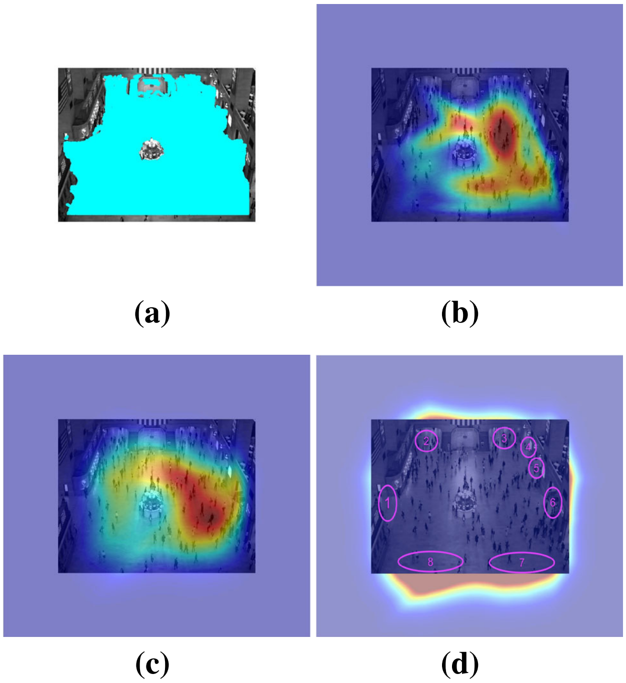


Fig. 16 Sample frames from TrainStation with a border added to the frame to show the **a** density of automatically generated people's locations, **b** fitted GMMs on the people's locations, **c** static rendezvous points and **d** rendezvous areas outside the field of view. The *numbered ellipses* indicate the entry/exit locations manually labeled by [71]

Interestingly, static rendezvous areas in Fig. 16c show where the rendezvous areas are almost complementary to the distribution of the people's locations and mostly concentrated close to the right-hand side exit location. By cropping the distribution of the rendezvous areas in the center of the scene, we can highlight the rendezvous areas outside the field of view (Fig. 16d). If we compare these areas with the entry/exit locations manually labeled by [71], we can observe that the approximations we obtain are

[8] Video results on the full sequence can be found here: ftp://motinas.elec.qmul.ac.uk/pub/ra_results/trainstation_border_a.zip ftp://motinas.elec.qmul.ac.uk/pub/ra_results/trainstation_border_b.zip.

very close to them. Locations 4 and 5 are not detected since they are located within the image, whereas rendezvous locations for 1 and 6 are slightly above the true location due to the perspective distortion of the view.

5.3 Computational complexity and execution time

We measure the computational complexity and execution time of the proposed algorithm by using a non-optimized Matlab code. The experiments are performed with more than 8,000 frames of TrainStation on a core Intel i5 2.4 GHz and 8Gb RAM. The execution time is measured per frame and it is shown in Fig. 17. The main constraint to be taken into account for the algorithmic real-time employment involves the operational computation of the rendezvous areas before events of interest occur. For example, in Fig. 7 we showed a particular application where 80-frame predictions can be computed.

At each frame t the main algorithmic steps that compute dynamic rendezvous areas once the buffer τ is full are: the computation of the displacement vectors, the extraction of the intersections of the straight lines (e.g., Gauss elimination method) and the unsupervised GMMs on the intersection points. Let s be the frame size, v the (maximum) velocity constraint used for the computation of the displacement vectors (e.g., optical flow) [39] and D the dimension of the observations (in our case 2). At t, I^t is the number of displacement vectors, J^t the number of intersections and M^t is the order of the mixture. The overall computational complexity of the proposed algorithm is $O(sv) + O({I^t}^2) + O(8J^tM^t)$ which corresponds to $O(\max(sv, {I^t}^2, 8J^tM^t))$. The complexity is time dependent since the crowd density over the video sequence can vary and the number of intersection points will vary accordingly.

The computational complexity for the calculation of the displacement vectors is linear in relation to the size of the image as there can be correlation-based algorithms that exploit the filter separability by using 1-D spatial search and gradient-based algorithms [38]. Hence, displacement vectors can be calculated with $O(sv)$ operations. Moreover, the optical flow is parallelizable and can be implemented in real time [7, 20, 39].

The extraction of the straight line intersections is quadratic as it is computed for each pair of displacement vectors. Note that the number of displacement vectors used in this case is usually much less than the number of motion vectors calculated with the optical flow as the motion vectors with a very small magnitude can be discarded. The execution time for the computation of the intersection points is provided as a function of the number of displacement vectors and increases non-linearly as the number of vectors increases (Fig. 17a). The parallelization is possible since the intersection points are calculated independently for each pair of straight lines and the operations can be distributed among different processors.

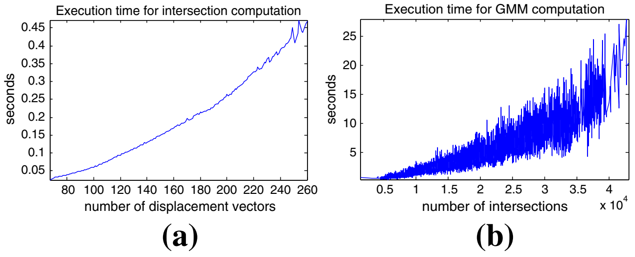


Fig. 17 Execution time calculated for the computation of **a** intersection points and **b** Gaussian Mixture Models. The oscillations in **b** are due to the distribution of the original data (intersection points) and the initialization of the clustering algorithm that is performed randomly

The complexity of an unsupervised GMM algorithm that uses E-M can be $O(4DJ^tM^t)$ [23]. The execution time for the GMM computation increases non-linearly with the number of intersections (Fig. 17b). The figure also shows oscillations that are due to two factors. The first factor involves the variation of the spatial distribution of the intersection points. In fact, when the intersection points do not follow Gaussian distributions, the algorithm may require more time to estimate the parameters of the mixture. The second factor involves the initialization of the fitting algorithm that is performed randomly. The fitting may require more execution time if the mean and covariance of the mixture are initialized with values far from those of the converging solution. This experimental evidence shows us that the processing stage for the GMM computation needs optimization, which can be addressed with real-time solutions proposed in literature such as [18, 32, 56]. For example, the E-M algorithm for GMMs can be parallelized. Kumar et al. [32] showed that estimating a mixture with 16 components and 76.8K observations, by using a GPU implementation on Geforce Quadro FX 5800 with 240 cores takes 42.4 ms. In our case, we observed that in a crowded environment there are no more than ten components per frame and 5K observations (intersection points).

6 Conclusions

We proposed a method to predict dynamic and static rendezvous areas. Dynamic rendezvous areas were analyzed over time, by evaluating the likelihood of the predicted areas with respect to the people's locations in the future. Static rendezvous areas were quantitatively validated by comparing the results with people's locations over time. The method utilizes displacement vectors to estimate the principal direction of people's motion. We use the intersections of prediction lines that provide us with the candidate rendezvous areas. Gaussian Mixture Models are further fitted on the intersection points for a continuous and smooth representation of the rendezvous areas. We also exploited the algorithm to infer the structure of the environment in locations outside the field of view of a camera by estimating rendezvous areas in unobservable locations. We presented a detailed survey of state-of-the-art methods for the recognition of human interactions and provided a list of publicly available datasets. Methods have been divided into those exploiting the global motion of the scene (e.g., optical flow) and those that use individual target patterns (e.g., multi-target tracking).

We hope that the proposed idea will inspire new research directions towards the prediction of interactions in videos. For example, a detailed analysis of the interpersonal distances that aims at predicting the type of interaction occurring according to proxemic theory.

References

1. Andriyenko, A., Schindler, K., Roth, S.: Discrete-continuous optimization for multi-target tracking. In: Proceedings of Computer Vision and Pattern Recognition, Providence. pp. 1926–1933 (2012)
2. Bazzani, L., Cristani, M., Murino, V.: Decentralized particle filter for joint individual-group tracking. In: Proceedings of Computer Vision and Pattern Recognition, Providence. pp. 1886–1893 (2012)
3. Benabbas, Y., Ihaddadene, N., Djeraba, C.: Motion pattern extraction and event detection for automatic visual surveillance. EURASIP **7**, 1 (2011)
4. Bera, A., Galoppo, N., Sharlet, D., Lake, A., Manocha, D.: Adapt: real-time adaptive pedestrian tracking for crowded scenes. In: Proceedings of Conference on Robotics and Automation, Hong Kong. (2014)
5. Borges, P., Conci, N., Cavallaro, A.: Video-based human behavior understanding: a survey. Trans. Circuits Syst. Video Technol. **23**(11), 1993–2008 (2013)
6. Bouman, C: Cluster: an unsupervised algorithm for modeling gaussian mixtures. http://engineering.purdue.edu/-bouman. (1998)
7. Bulthoff, H., Little, J., Poggio, T.: A parallel algorithm for real-time computation of optical flow. Nature **337**(6207), 549–553 (1989)
8. Calderara, S., Cucchiara, R.: Understanding dyadic interactions applying proxemic theory on videosurveillance trajectories. In: Proceedings of Computer Vision and Pattern Recognition Workshop, Providence. pp. 20–27 (2012)
9. Chang, M.C., Krahnstoever, N., Ge, W.: Probabilistic group-level motion analysis and scenario recognition. In: Proceedings of International Conference on Computer Vision, Barcelona. pp. 747–754 (2011)
10. Chaquet, J., Carmona, E., Fernandez-Caballero, A.: A survey of video datasets for human action and activity recognition. Comput. Vis. Image Underst. **117**(6), 633–659 (2013)
11. Chen, D.Y., Huang, P.C.: Motion-based unusual event detection in human crowds. J. Vis. Commun. Image R. **22**(2), 178–186 (2011)
12. Chen, F., Cavallaro, A.: Detecting group interactions by online association of trajectory data. In: Proceedings of Acoustics, Speech, and Signal Processing, Vancouver. pp. 1754–1758 (2013)
13. Cong, Y., Liu, J.Y.J.: Sparse reconstruction cost for abnormal event detection. In: Proceedings of Computer Vision and Pattern Recognition, Colorado Springs. pp. 3449–3456 (2011)
14. Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Bue, A.D., Menegaz, G., Murino, V. : Social interaction discovery by statistical analysis of *F*-formations. In: Proceedings of British Machine Vision Conference, Dundee. pp. 1–12 (2011a)
15. Cristani, M., Paggetti, G., Vinciarelli, A., Bazzani, L., Menegaz, G., Murino, V.: Towards computational proxemics: inferring social relations from interpersonal distances. In: Proceedings of Internation Conference on Social Computing, Sydney. pp. 290–297 (2011b)
16. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for human detection. In: Proceedings of Computer Vision and Pattern Recognition, San Diego. pp. 886–893 (2005)
17. Farenzena, M., Tavano, A., Bazzani, L., Tosato, D., Paggetti, G., Menegaz, G., Murino, V., Cristani, M.: Social interactions by visual focus of attention in a three-dimensional environment. In: Workshop on Pattern Recognition and Artificial Intelligence for Human Behaviour Analysis, Reggio Emilia. (2009)
18. Fassold, H., Rosner, J., Schallauer, P., Bailer, W.: Realtime KLT feature point tracking for high definition video. In: Proceedings of Computer Graphics, Computer Vision and Mathematics, Plzen. pp. 40–47 (2009)
19. Fathi, A., Hodgins, J., Rehg, J.: Social interactions: a first-person perspective. In: Proceedings of Computer Vision and Pattern Recognition, Providence. pp. 1226–1233 (2012)
20. Garcia-Rodriguez, J., Orts-Escolano, S., Angelopoulou, A., Psarrou, A., Azorin-Lopez, J., Garcia-Chamizo, J.: Real time motion estimation using a neural architecture implemented on GPUs. J. Real-Time Image Process. (2014)
21. Granger, C.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica **37**(3), 424–438 (1969)
22. Greggio, N., Bernardino, A., Laschi, C., Dario, P., Santos-Victor, J.: Self-adaptive Gaussian mixture models for real-time video segmentation and background subtraction. In: Proceedings of Intelligent Systems Design and Applications, Cairo. pp. 983–989 (2010)
23. Greggio, N., Bernardino, A., Laschi, C., Dario, P., Santos-Victor, J.: Fast estimation of Gaussian mixture models for image segmentation. Mach. Vis. Appl. **23**(4), 773–789 (2012)
24. Hall, E.: The Hidden Dimension: Handbook for Proxemic Research. Anchor Books Doubleday, New York (1966)
25. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Phys. Rev. E **51**(5), 4282–4286 (1995)
26. Jin, B., Hu, W., Wang, H.: Human interaction recognition based on transformation of spatial semantics. IEEE Sign. Process. Lett. **19**(3), 139–142 (2012)
27. Kendon, A.: Studies in the Behavior of Social Interaction. Indiana Univeristy Press, Bloomington (1977)
28. Kendon, A.: Development of Multimodal Interfaces: Active Listening and Synchrony. Spacing and Orientation in Co-present, Interaction, pp. 1–15. Springer, Berlin (2009)
29. Kim, K., Grundmann, M., Shamir, A., Matthews, I., Hodgins, J., Essa, I.: Motion field to predict play evolution in dynamic sport scenes. In: Proceedings of Computer Vision and Pattern Recognition, San Francisco. pp. 840–847 (2010)
30. Kirby, R.: Social Robot Navigation. Ph.D. Thesis (CMU-RI-TR-10-13), Robotics Institute, Carnegie Mellon University, Pittsburgh (2010)

31. Krausz, B., Bauckhage, C.: Loveparade 2010: automatic video analysis of a crowd disaster. Comput. Vis. Image Underst. **116**(3), 307–319 (2012)
32. Kumar, N., Satoor, S., Buck, I.: Fast parallel expectation maximization for Gaussian mixture models on GPUs using CUDA. In: Proceedings of High Performance Computing and Communications, Seoul. pp. 103–109 (2009)
33. Lan, T., Sigal, L., Mori, G.: Social roles in hierarchical models for human activity recognition. In: Proceedings of Computer Vision and Pattern Recognition, Providence. pp. 1354–1361 (2012)
34. Laptev, I.: On space–time interest points. Intern. J. Comput. Vis. **64**(2/3), 107–123 (2005)
35. Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. Intern. J. Comput. Vis. **77**(1), 259–289 (2008)
36. Lester, P.M.: Visual Communication: Images with Messages. Wadsworth Publishing Co Inc., Belmont (2002)
37. Li, R., Porfilio, P., Zickler, T.: Finding group interactions in social clutter. In: Proceedings of Computer Vision and Pattern Recognition, Columbus. pp. 2722–2729 (2013)
38. Liu, H., Hong, T.H., Herman, M., Chellappa, R.: Accuracy vs. efficiency trade-offs in optical flow algorithms. Comput. Vis. Image Underst. **72**(3), 271–286 (1996)
39. Liu, H., Hong, T.H., Herman, M., Chellappa, R.: A general motion model and spatio-temporal filters for computing optical flow. Intern. J. Comput Vis. **22**(2), 141–172 (1997)
40. Liu, J., Carr, P., Collins, R., Liu, Y.: Tracking sports players with context-conditioned motion models. In: Proceedings of Computer Vision and Pattern Recognition, Portland. pp. 1830–1837 (2013)
41. Lowe, D.: Object recognition from local scale-inveriant feature. In: Proceedings of International Conference on Computer Vision, Corfu. pp. 1150–1157 (1999)
42. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of International Joint Conference on Artificial Intelligence, San Francisco. pp. 674–679 (1981)
43. Mazzon, R., Poiesi, F., Cavallaro, A.: Detection and tracking of groups in crowd. In: Proceedings of Advanced Video and Signal Based Surveillance, Krakow. pp. 202–207 (2013)
44. McKenna, S., Nait-Charif, H.: Learning spatial context from tracking using penalised likelihoods. In: Proceedings of International Conference on Pattern Recognition, Cambridge. pp. 138–141 (2004)
45. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: Proceedings of Computer Vision and Pattern Recognition, Miami. pp. 935–942 (2009)
46. Mehran, R., Moore, B., Shah, M.: A streakline representation of flow in crowded scenes. In: Proceedings of European Conference in Computer Vision, Crete. pp. 439–452 (2010)
47. Nayak, N., Zhu, Y., Roy-Chowdhury, A.: Vector field analysis for multi-object behavior modeling. Comput. Vis. Image Underst. **31**(6–7), 460–472 (2013)
48. Needleman, S., Wunsch, C.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**(3), 443–453 (1970)
49. Oliver, N.: Towards perceptual intelligence: statistical modeling of human individual and interactive behaviors. Ph.D. thesis, Massachusetts Institute Technology (MIT), Media Lab, Cambridge, Mass (2000)
50. Papadakis, P., Spalanzani, A., Laugier, C.: Social mapping of human-populated environments by implicit function learning. In: Proceedings of Intelligent Robots and Systems, Tokyo. pp. 1701–1706 (2013)
51. Pellegrini, S., Ess, A., Schindler, K., Gool, L.V.: You will never walk alone: modeling social behavior for multi-target tracking. In: Proceedings of Internation Conference on Computer Vision, Kyoto. pp. 261–268 (2009)
52. Pellegrini, S., Ess, A., Gool, L.V.: Improving data association by joint modeling of pedestrian trajectories and groupings. In: Proceedings of European Conference on Computer Vision, Heraklion, Crete. pp. 452–465 (2010)
53. Poiesi, F., Danyial, F., Cavallaro, A.: Detector-less ball localization using context and motion flow analysis. In: Proceedings of International Conference on Image Processing, Hong Kong. pp. 3913–3916 (2010)
54. Raghavendra, R., Bue, A.D., Cristani, M., Murino, V.: Optimizing interaction force for global anomaly detection in crowded scenes. In: Proceedings of Internation Conference on Computer Vision Workshop, Barcelona. pp. 136–143 (2011)
55. Ryoo, M., Aggarwal, J.L.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: Proceedings of International Conference on Computer Vision, Kyoto. pp. 1593–1600 (2009)
56. Salvadori, C., Petracca, M., del Rincon, J.M., Velastin, S.A., Makris, D.: An optimisation of Gaussian Mixture Models for integer processing units. J. Real-Time Image Process (2014).
57. Sfikas, G., Constantinopoulos, C., Likas, A., Galatsanos, N.P.: An analytic distance metric for Gaussian mixture models with application in image retrieval. Artif. Neural Netw. **3697**, 835–840 (2005)
58. Sinha, S., Frahm, J.M., Pollefeys, M., Genc, Y.: GPU-based video feature tracking and matching. Technical Report TR 06–012, Department of Computer Science, UNC Chapel Hill, Chapel Hill (2006)
59. Sochman, J., Hogg, D.: Who knows who inverting the social force model for finding groups. In: Proceedings of International Conference on Computer Vision Workshop, Barcelona. pp. 830–837 (2011)
60. Soldera, F., Calderara, S., Cucchiara, R.: Structured learning for detection of social groups in crowd. In: Proceedings of Advanced Video and Signal Based Surveillance, Krakow. pp. 7–12 (2013)
61. Solmaz, B., Moore, B., Shah, M.: Identifying behaviors in crowd scenes using analysis for dynamical systems. IEEE Trans. PAMI **34**(10), 2064–2070 (2012)
62. Su, H., Yang, H., Zheng, S., Fan, Y., Wei, S.: The large-scale crowd behavior perception based on spatio-temporal viscous fluid fields. IEEE Trans. Info. Forens. Sec. **8**(10), 1556–1589 (2013)
63. Suk, H.I., Jain, A., Lee, S.W.: A network of dynamic probabilistic models for human interaction analysis. IEEE Trans. Circuits Syst. Video Technol. **21**, 932–945 (2011)
64. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: Proceedings of Computer Vision and Pattern Recognition, San Francisco. pp. 2432–2439 (2010)
65. Taj, M., Cavallaro, A.: Recognizing Interactions in Video. Intelligent Multimedia Analysis for Security Applications, vol. 282/2010. Springer, Berlin (2010)
66. Taj, M., Cavallaro, A.: Interaction recognition in wide areas using audiovisual sensors. In: Proceedings of Internation Conference on Image Processing, Orlando. pp. 1113–1116 (2012)
67. Tao, J., Klette, R.: Integrated pedestrian and direction classification using a random decision forest. In: Proceedings of International Conference on Computer Vision Workshop, Sydney. pp. 230–237 (2013)
68. Wang, X., Ma, X., Grimson, W.: Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian model. IEEE Trans. Patt. Anal. Mach. Intell. **31**(3), 539–555 (2009)
69. Zanotto, M., Cristani, L.B.B., Murino, V.: Online bayesian nonparametrics for group detection. In: Proceedings of British Machine Vision Conference, Surrey. pp. 111.1–111.12 (2012)
70. Zhao, M., Turner, S., Cai, W.: A data-driven crowd simulation model based on clustering and classification. In: Proceedings of

Distributed Simulation and Real Time Applications, Delft. pp. 125–134 (2013)

71. Zhou, B., Wang, X., Tang, X.: Understanding collective crowd behaviors: learning a mixture model of dynamic pedestrian-agents. In: Proceedings of Computer Vision and Pattern Recognition, Providence. pp. 2871–2878 (2012)
72. Zhou, B., Tang, X., Wang, X.: Measuring the collectiveness. In: Proceedings of Computer Vision and Pattern Recognition, Columbus. pp. 3049–3056 (2013)

Fabio Poiesi received his Ph.D. in Electronic Engineering and Computer Science from the Queen Mary University of London (UK) in 2014. He received B.Sc. and M.Sc. degrees in Telecommunication Engineering from the University of Brescia (Italy) in 2007 and 2010, respectively. His research field involves multi-target tracking in highly populated scenes, behavior understanding and performance evaluation of tracking algorithms. Dr. Poiesi is currently working as a researcher at Queen Mary University of London under the European project COPCAMS (copcams.eu).

Andrea Cavallaro is Professor of Multimedia Signal Processing and Director of the Centre for Intelligent Sensing at Queen Mary University of London, UK. He received his Ph.D. in Electrical Engineering from the Swiss Federal Institute of Technology (EPFL), Lausanne, in 2002. He was a Research Fellow with British Telecommunications (BT) in 2004/2005 and was awarded the Royal Academy of Engineering teaching Prize in 2007; three student paper awards on target tracking and perceptually sensitive coding at IEEE ICASSP in 2005, 2007 and 2009; and the best paper award at IEEE AVSS 2009. Prof. Cavallaro is Area Editor for the IEEE Signal Processing Magazine and Associate Editor for the IEEE Transactions on Image Processing. He is an elected member of the IEEE Signal Processing Society, Image, Video, and Multidimensional Signal Processing Technical Committee, and chair of its Awards committee. He served as an elected member of the IEEE Signal Processing Society, Multimedia Signal Processing Technical Committee, as Associate Editor for the IEEE Transactions on Multimedia and the IEEE Transactions on Signal Processing, and as Guest Editor for seven international journals. He was General Chair for IEEE/ACM ICDSC 2009, BMVC 2009, M2SFA2 2008, SSPE 2007, and IEEE AVSS 2007. Prof. Cavallaro was Technical Program chair of IEEE AVSS 2011, the European Signal Processing Conference (EUSIPCO 2008) and of WIAMIS 2010. He has published more than 130 journal and conference papers, one monograph on Video tracking (2011, Wiley) and three edited books: Multi-camera networks (2009, Elsevier); Analysis, retrieval and delivery of multimedia content (2012, Springer); and Intelligent multimedia surveillance (2013, Springer).