Detection and tracking of groups in crowd

Riccardo Mazzon, Fabio Poiesi, Andrea Cavallaro
Centre for Intelligent Sensing, Queen Mary University of London
Mile End Road, London E1 4NS, UK
{riccardo.mazzon,fabio.poiesi,andrea.cavallaro}@eecs.qmul.ac.uk

Abstract

We propose a method to detect and track interacting people by employing a framework based on a Social Force Model (SFM). The method embeds plausible human behaviors to predict interactions in a crowd by iteratively minimizing the error between predictions and measurements. We model people approaching a group and restrict the group formation based on the relative velocity of candidate group members. The detected groups are then tracked by linking their interaction centers over time using a buffered graph-based tracker. We show how the proposed framework outperforms existing group localization techniques on three publicly available datasets, with improvements of up to 13% on group detection.

1. Introduction

About 50-70% of human walking activity takes place in groups [11]: video monitoring of spatially interacting humans is therefore very important for analyzing people's behaviors [3, 10]. The automatic localization of groups is very challenging in crowded scenarios, such as public squares or large malls, where spatial proximity alone does not help to determine whether or not people are interacting. Groups can be detected online, offline or with a delay.

Online methods enable the localization of interactions without using future information about the dynamics in the scene [1, 12]. For example, a Decentralized Particle Filtering (DPF) for group detection and tracking can be used where the states of the filter contain position and velocity information of people, and labels of the group affiliation of each person. Furthermore, an unsupervised method for group detection based on Dirichlet Process Mixture Model (DPMM) can also be employed [12], where motion patterns along with social constraints based on rules of proxemics are used to determine group formations. *Offline* methods process the information extracted from the whole video in a batch. The extracted human motion patterns (e.g. positions and velocity) are temporally analyzed in order to determine the affiliation of each subject to a particular group and group detection is performed on the overall permanence of people within a group. Common directions and velocities of humans processed with a bottom-up hierarchical clustering [2, 3] or an optimization based on Lagrangian theory [9] can be employed to model group behaviors. Alternative techniques are based on *social forces* that analyze relative motion patterns [5, 6, 11]. These approaches aim to recognize groups that contain people who know each other, rather than localizing short interactions. *Delayed* methods can use the SFM for group modeling, whereas the final group decision [11] is taken with an offline error minimization process. Interestingly, this algorithm outperforms state-of-the-art approaches in several difficult scenarios even if the group modeling only analyzes people's movements and does not employ explicit human behavior constraints for group formation. However, the approach struggles to detect instantaneous interactions in some simple situations which can be partially address by offline approaches [3].

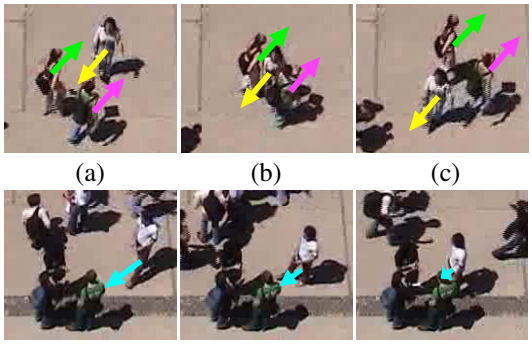


Figure 1. Visualization of plausible human behavior constraints: (a-c) person crossing a group (two people) not detected as part of the group; (d-f) person approaching a stationary group detected as a member only if he/she decelerates and stops in its proximity.

In this paper, we extend the delayed SFM for group detection [11] by defining plausible human behaviors for the localization of group formations (Fig. 1). We improve the model to enable group detection in situations that were not previously possible by incorporating relationships such as walking in the same direction and decelerating when ap-

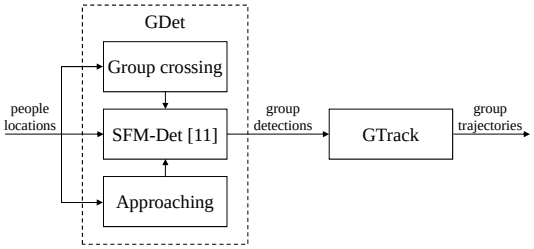


Figure 2. Block diagram of the proposed approach.

proaching individuals who are standing still. Moreover, we track each group over time with a graph-based tracker to enforce the spatio-temporal consistency of the detections. We validate the proposed framework on three different datasets and show that it outperforms existing methods using the Group Detection Success Rate (GDSR) [1].

The paper is organized as follows. Section 2 describes the proposed method for group detection and tracking. The method is validated and compared with existing approaches in Sec. 3. Finally, Sec. 4 draws conclusions and outlines future directions for research.

2. Interaction localization

In this section, we describe the solution for detection (Sec. 2.1) and tracking (Sec. 2.2) of groups. The former models people interacting with each other and the latter links the detected centers of interactions over time. Figure 2 shows the block diagram of the overall solution.

2.1. Interaction detection

Let $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ be the set of n people in the monitored scene, and $\mathbf{p}_i^t = (x_i^t, y_i^t)$ be the feet position of person p_i at time t on the rectified image. The SFM (SFM-Det) [11] has been used to detect interactions among people by analyzing the instantaneous forces; however, this method fails to detect groups in situations when the paths of walking people cross over each other or when people pass near a stationary group. In our approach, we specifically model these situations.

In the SFM-Det, the *desired velocity* of a person is defined as the average velocity observed in the time interval $[t, t+s]$, and the force $\mathbf{f}_{D_i}^t$ as the displacement between the actual velocity and the desired velocity. Let us define the set of people without p_i to be $\mathcal{H}_i = \mathcal{P} \setminus p_i$ and the contour of non-walkable areas to be W (walls/barriers), where W is the set of spatial locations defined by (x_W, y_W) that are the horizontal and vertical coordinates, respectively. The sums of repulsive forces, $\mathbf{f}_{\mathcal{H}_i}^t$ and $\mathbf{f}_{W_i}^t$, are inversely proportional to the distance between $\mathbf{p}_i^t$ and $\mathcal{H}_i$, and $\mathbf{p}_i^t$ and W in order for p_i to be at a comfortable distance from other people $\mathcal{H}_i$ and from walls/barriers W, respectively. The attractive force, $\mathbf{f}_{G_{i,g}}^t$, is the force that keeps people within the same group $G_{i,g} \subseteq \mathcal{P}$ [11], where g is the unique group index (in the rest of the paper, we omit g for better readability). The sum of forces describing the variation of people's movement is

$$m_i \frac{d\mathbf{v}_i^t}{dt} = \mathbf{f}_{D_i}^t + \mathbf{f}_{\mathcal{H}_i}^t + \mathbf{f}_{W_i}^t + \mathbf{f}_{G_i}^t, \quad (1)$$

where $\frac{d\mathbf{v}_i^t}{dt}$ is the actual variation of the velocity over time and m_i the mass of people. The left-hand side of Eq. 1 models the variation in space of a person's movement at each time instant and the position of each person is then predicted using

$$\mathbf{p}_i^{*t+s} = \mathbf{p}_i^t + s\Big(\frac{d\mathbf{v}_i^t}{dt}\tau + \overline{\mathbf{v}}_i^t\Big), \quad (2)$$

where $\mathbf{p}_i^{*t+s}$ indicates the expected position of p_i after time $t+s$ determined using the SFM, τ is the time interval during which the Eq. 1 is calculated, and $\overline{\mathbf{v}}_i^t$ is the *actual velocity* obtained as the average velocity within $[t-s, t]$. Lower values of s would result in a model that is sensitive to noise, while higher values of s could not reliably describe abrupt velocity variations. We set $s = 10$ as in [11].

Compared to the original SFM [4], the term $\mathbf{f}_{G_i}^t$ (Eq. 1) is introduced for group behavior modeling [5, 6] that is used for interaction detection in an iterative algorithm [11]. This algorithm evaluates whether the SFM prediction (Eq. 2) commits more error, δ, when there are no group forces involved ($\mathbf{f}_{G_i}^t = (0,0)$), or when there is an active group force keeping people together and inhibiting them from being repulsed by each other. In [11], all people can interact with each other (within a certain gating radius) in order to form a group, thus enabling the group detection only when people do not walk or stand next to each other. However, in crowded cases where people can cross existing groups and walk very close to other people but without stopping (Fig. 3), SFM-Det may fail. In order to address these challenging situations, we propose to restrict the set of potentially interacting people to those walking in similar directions (*group crossing*) and to those decelerating when approaching a stationary group (*approaching*). The former model defines that people walking in opposite directions are non-interacting, while the latter restricts people interacting with a stationary group to those decelerating in its proximity and almost stopping. We define our method as GDet. Let us call $\overline{\mathcal{H}}_i^t \subseteq \mathcal{H}_i$ and $\widehat{\mathcal{H}}_i^t \subseteq \mathcal{H}_i$ the sets of people selected by group crossing and approaching, respectively, for person p_i at time t. Equation 1 and Eq. 2 become, respectively

$$m_i \frac{d\mathbf{v}_i^{*t}}{dt} = \mathbf{f}_{D_i}^t + \mathbf{f}_{\mathcal{H}_i^*}^t + \mathbf{f}_{W_i}^t + \mathbf{f}_{G_i}^t, \quad (3)$$

$$\overline{\mathbf{p}}_i^{*t+s} = \mathbf{p}_i^t + s\Big(\frac{d\mathbf{v}_i^{*t}}{dt}\tau + \overline{\mathbf{v}}_i^t\Big), \quad (4)$$

where the repulsive force, $\mathbf{f}_{\mathcal{H}_i^*}^t$, acts between p_i and $\mathcal{H}_i^{*t} = \overline{\mathcal{H}}_i^t \cup \widehat{\mathcal{H}}_i^t$ ($\mathcal{H}_i^{*t} \subseteq \mathcal{H}_i$), and $d\mathbf{v}_i^{*t}/dt$ and $\overline{\mathbf{p}}_i^{*t+s}$ are the new actual variation of the velocity and the new prediction of person p_i position, respectively. Note that in Eq. 3, the group force $\mathbf{f}_{G_i}^t$ is only applied between p_i and $\mathcal{H}_i^{*t}$.

Group crossing. Let us consider two people, p_1 and p_2, interacting with each other or with other people, and walking in the same direction within a range of $180°$. If a third person p_3 walks close to p_1 and p_2 in the opposite direction, the movement of p_3 biases the group detection of SFM-Det [11] that erroneously classifies p_1 and p_2 as non-interacting (Fig. 3(a-c)). This problem can be addressed by explicitly modeling the group crossing that allows people interacting with other subjects only if their direction of motion is coherent within a range of $180°$. Let us define $\mathbf{f}_{ij}^t$ the (repulsive) interaction force generated by p_j on p_i, where $p_j \in \mathcal{H}_i$. Initially, two groups of people (p_1 with p_2, and p_4 with p_5) walk in the same direction and far apart enough to not be detected as a single group, while a fifth person (p_3) walks in the opposite direction. With SFM-Det, when p_3 is nearby p_2 and p_4, the repulsive forces $\mathbf{f}_{23}^t$ and $\mathbf{f}_{43}^t$ act on their masses, and when people are almost aligned, they cancel out with $\mathbf{f}_{21}^t$ and $\mathbf{f}_{45}^t$, respectively, thus obtaining

$$\begin{aligned} \mathbf{f}_{21}^t &\cong -\mathbf{f}_{23}^t \\ \mathbf{f}_{45}^t &\cong -\mathbf{f}_{43}^t, \end{aligned} \tag{5}$$

a common situation in crowded scenarios (Fig. 1). Accepted predictions for the movements of p_2 and p_4 (Eq. 2) are obtained without any group forces, thus leading to a missed detection of the two groups (groups are detected only in the presence of a group force). With the inclusion of the group crossing constraint, p_3 does not influence p_2 and p_4 movements with a repulsive force since their motion direction is opposite. Group forces between p_1 and p_2, and p_4 and p_5 are generated, Eq. 4 provides accepted predictions and GDet correctly detects the groups. Figure 4(a) reports a schematic representation of the forces.

Approaching. When dealing with crowded scenes where frequent meetings occur, a single person p_i may interact with a stationary group $\mathcal{S} \subseteq \mathcal{H}_i$, or may pass very close to it in order to shorten the path to reach his/her goal. The vicinity of p_i to $\mathcal{S}$ generates a set of high repulsive forces, $\mathbf{f}_{\mathcal{S}_i}$, that in SFM-Det [11] results in spurious group detections. In order to address this problem, we restrict the possible people interacting with $\mathcal{S}$ to only those decelerating in proximity of $\mathcal{S}$ and stopping within $\widehat{t}$ frames. Figure 4(b) shows the SFM forces with and without considering the approaching model. When p_3 is close to p_1 and p_2, SFM-Det allows the repulsive forces generated by p_3 on p_1 ($\mathbf{f}_{13}^t$) and on p_2 ($\mathbf{f}_{23}^t$) to influence their movement predictions (Eq. 2). In order to have acceptable predictions, $\mathbf{f}_{13}^t$ and $\mathbf{f}_{23}^t$ have to be balanced by group forces that make p_1, p_2 and p_3 being detected as in the same group, thus resulting in a false positive group detection for p_3. However, this problem is solved by including the approaching constraint where the interaction of p_3 with p_1 and p_2 is not allowed unless p_3 decelerates in proximity of p_1 and p_2, as if a meeting was about to happen. Figure 3(d-f) shows an example of group detection using the approaching model.

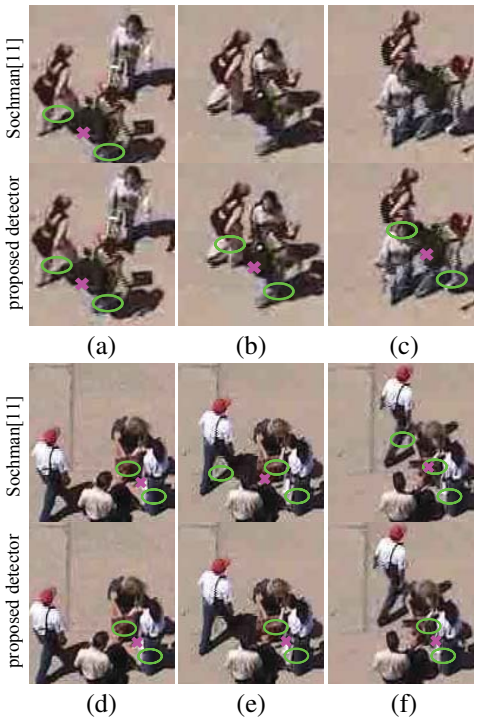


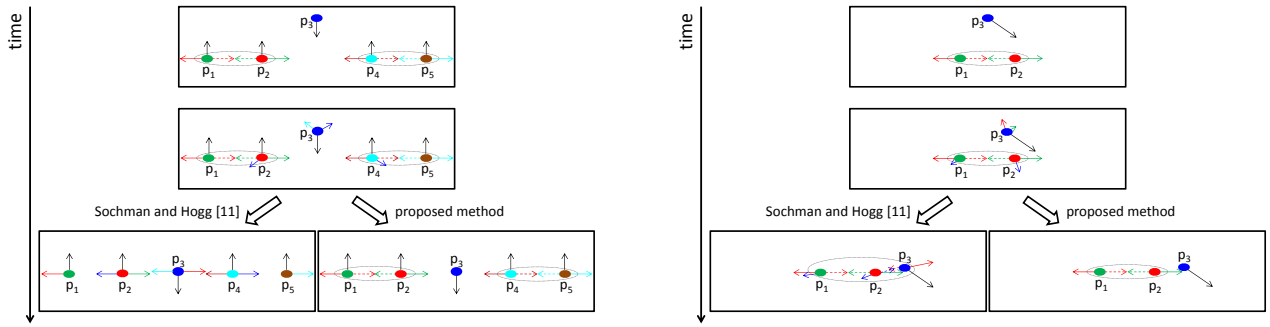
Figure 3. Sample group detection results in the case of a person (a-c) crossing and (d-f) passing nearby a group.

2.2. Group tracking

After frame-by-frame interaction localization is performed (GDet), we enforce the temporal consistency of the group detections (interaction centers) using a tracking framework. We define the interaction centers as the centroids of the positions of the people that form each group and we employ a buffered greedy graph-based multi-target tracker [8] (GTrack). At first, short tracks are generated by associating consecutive centroids with Hungarian algorithm (HA)[1]. The association cost used by HA is calculated with the ℓ-2 norm on the 2D centroid positions. Longer tracks are subsequently extracted using GTrack that pairwise matches short tracks until no alternative better pairings are found. GTrack determines the affinities among the short tracks using position and velocity information, and the association process is performed within a short temporal buffer and involves a sliding window of Θ frames overlapping for θ frames. When short tracks are associated, the missing centroids between the last location of an earlier short track and the initial location of a later one are generated by 2D interpolation. The people forming the group of the earlier short track are propagated up to the later short track.

In Fig. 5, we can see the effectiveness of GTrack on the interaction centers. A group of two people (light-blue track under white arrow) is passing nearby another group (brown) and, initially, the detector correctly localizes the interaction

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



(a) (b)

Figure 4. Difference between the proposed approach and [11] in how the forces act on people (a) while crossing a group and (b) approaching. Colored ellipses represent people ($\mathcal{P} = \{p_i\}_{i=1}^{5}$); the black arrow is the vector of movement ($\mathbf{f}_{D_i}^t$); solid and colored arrows are the repulsive forces ($\mathbf{f}_{\mathcal{H}_i}^t$); dotted arrows are the attractive forces that form the group ($\mathbf{f}_{G_i}^t$); and the dotted ellipses are the detected groups.

Table 1. Details of the datasets used in the experiments. Key - ppg: people per group.

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

centers (groups) (Fig. 5(a)). When the light-blue group is closer to the brown group (Fig. 5(b,c)) the detector fails and assigns the people of the light-blue group to the brown group. However, the tracker manages to recover this erroneous assignment (Fig. 5(d)) and returns to tracking the two people belonging to the light-blue group.

3. Results

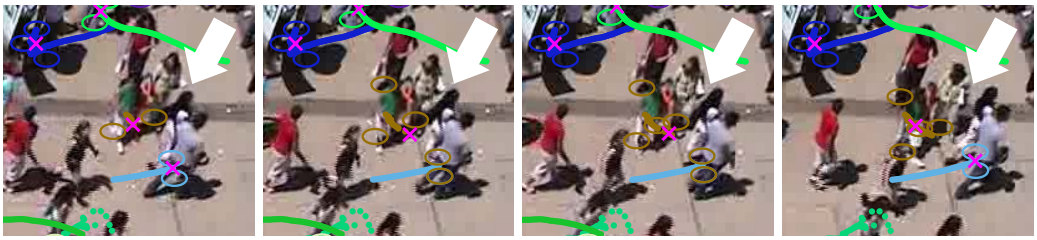
In order to validate the method for localizing groups, we compare it with methods [1, 11, 12] using ETH [7], HOTEL [7], and Student003 [11] datasets. ETH contains people mainly entering/exiting a building; HOTEL has more complex people's movement because of the presence of a tram stop with various barriers; Student003 presents a challenging scenario where people walk in unpredictable directions and get close to each other. In these datasets, different types of groups are formed, ranging from those in motion to those standing still. Table 1 provides the details of each dataset. The Joint Individual-Group Tracking (JIGT) [1] is based on an online DPF and is characterized by two conditionally dependent subspaces that model people's motion and group formations, respectively. Instead, in [12], observations of people's locations and velocities are generated using a tracker, and group detection is performed online by modeling groups as infinite mixtures solved using DPMM.

For group detection (GDet), we use the same parameter setting of [11], except for $\delta = 3$ (instead of $\delta = 0$) in order to remove noisy group detections. A person with average speed less than their shoulder radius (0.35 $meters$ [11]) within a one-second window is considered to be standing still and $\widehat{t} = s$ for the approaching model. For group tracking (GTrack), we set $\Theta = 25$ and $\theta = 5$ frames. Like [1, 11, 12], we consider the single-camera person tracking task solved. We compare the results of GDet and the algorithm of [11] without the offline decision (SFM-Det), and those of GTrack applied to the output of GDet and of SFM-Det (let us call them GTrack and SFM-TR, respectively). For JIGT [1] and DPMM [12], we provide the results from the related papers. The evaluation is performed with the mean of all frames of the GDSR proposed in [1], which calculates the rate of correctly detected groups. A correct detection of a group contains at least 60% of its members [1]. Table 2 reports the quantitative evaluation (the results for JIGT and DPMM are only available for the ETH dataset). The performance of GDet and GTrack is superior than that of other methods, thus proving the effectiveness of the proposed model for group localization (Sec. 2). The results of GDet compared to SFM-Det in HOTEL and Student003 are dramatically improved (11% and 13%, respectively) since the scene contains people standing still and groups are formed next to each other, whereas in ETH the improvement is minor (1%) because groups are relatively far from each other. Moreover, an improvement of 15% compared to DPMM is obtained in ETH, probably because this work is designed for detecting people switching groups.

Table 2. Result comparison using GDSR [1]. Key - GDet: proposed group detection; GTrack: proposed group tracking on the output of GDet; SFM-Det: group detection of [11]; SFM-TR: proposed group tracking on the output of SFM-Det; JIGT: group detection and tracking of [1]; DPMM: group detection of [12].

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

The group tracker (GTrack) improves the performance of GDet by 2% and 1% in ETH and Student003, respectively, whereas in HOTEL the results remain the same (89%). This is due to the high performance of GDet in HOTEL where



(a) (b) (c) (d)

Figure 5. Example of track recovery of a two-person group (light-blue track). Colored circle: affiliation of the people to a group; Colored line: trajectory; Magenta cross: group detection. (a) Light-blue group is correctly detected and tracked; (b,c) for some subsequent frames the light-blue group is erroneously detected as part of a neighboring group; (d) the light-blue group is correctly recovered by the tracker.

groups are constantly detected over time, unlike in ETH and Student003 where GDet provides less consistent input to the tracker that can then link the interaction centers over time. GTrack also improves the performance of SFM-TR by 1%, 3% and 2% in the three dataset and it has a 24% improvement compared to JIGT in ETH because, like DPMM, this work is more suitable to detect switchings of groups. Figure 6 shows the qualitative results as comparison between GTrack and SFM-TR. We can see some of the challenging situations where the proposed modeling is effective, and how most of the groups in the scene are linked over time and correctly tracked. GTrack shows better performance than SFM-TR, for example in Fig. 6(d) where a stationary group is correctly detected and tracked, even when another one passes close to it. Likewise, in Fig. 6(g) on the left, two groups (light blue and purple) that cross each other are consistently localized. In some situations, the group localization may fail. For example in Fig. 6(h), the three people in the middle of the frame are not localized as part of the same group because they just joined together and their group formation is highly unstable, that is they keep moving apart and joining back together, while walking and passing through other groups of people. However, SFM-TR can correctly localize part of this group (two out of three people), even for only few frames until when the group crosses another one coming from the opposite direction.

4. Conclusions

This paper proposed a detection algorithm for interacting people and a tracking algorithm for linking interaction centers over time. The group detector improved the state-of-the-art approach [11] by using walking direction and velocity variation constraints. The walking direction is measured for each person and group formations are permitted only if neighboring people share the same direction. Group formation is also modeled for the cases when a person engaging a group of isolated people decreases his/her velocity. The temporal consistency of the localization results were then improved by a buffered graph-based tracker. We finally showed that our framework outperforms state-of-the-art methods. As future work, the proposed group detection and tracking can be further improved using online tracking approaches in order to be applicable in time-critical applications.

References

[1] L. Bazzani, M. Cristani, and V. Murino. Decentralized particle filter for joint individual-group tracking. In *Proc. of IEEE CVPR*, 2012.

[2] W. Ge, R. T. Collins, and B. Ruback. Automatically detecting the small group structure of a crowd. In *Proc. of IEEE WACV*, 2009.

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

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

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

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

[7] S. Pellegrini, A. Ess, and L. V. Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *Proc. of ECCV*, 2010.

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

[9] Z. Qin. Improving multi-target tracking via social grouping. In *Proc. of IEEE CVPR*, 2012.

[10] B. Solmaz, B. Moore, and M. Shah. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *IEEE Tran. on PAMI*, 34(10):2064–2070, Oct. 2012.

[11] J. Šochman and D. C. Hogg. Who knows who - inverting the social force model for finding groups. In *Proc. of IEEE ICCVW*, 2011.

[12] M. Zanotto, L. Bazzani, M. Cristani, and V. Murino. Online bayesian nonparametrics for group detection. In *Proc. of BMVC*, 2012.

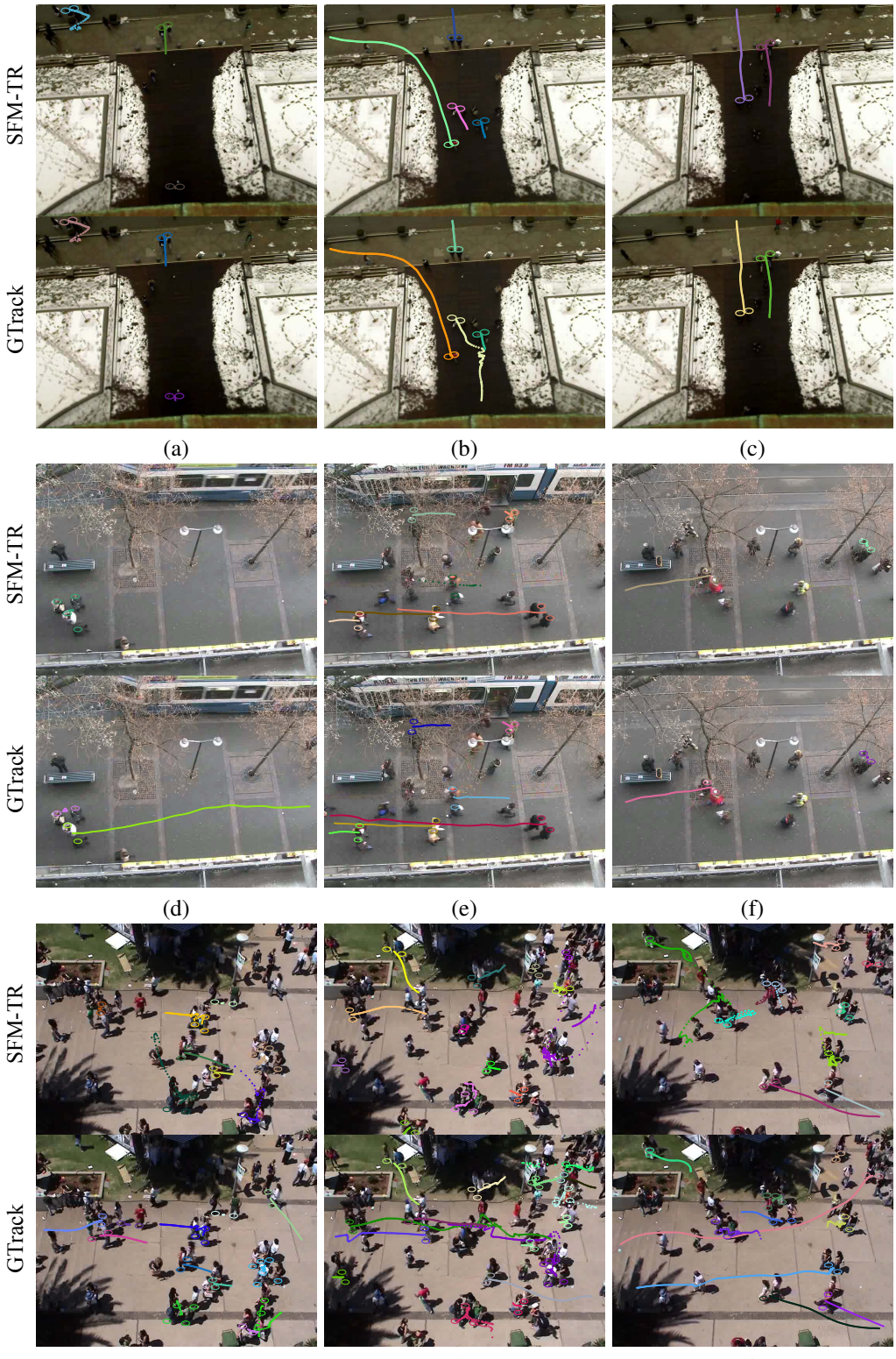


Figure 6. Sample group tracking results obtained with SFM-TR and GTrack on the (a-c) ETH, (d-f) HOTEL, and (g-i) Student003 datasets.