Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings

Stefano Pellegrini[1], Andreas Ess[1], and Luc Van Gool[1,2]

[1] Computer Vision Laboratory
ETH Zurich
[2] ESAT-PSI / IBBT
KU Leuven
{stefpell,aess,vangool}@vision.ee.ethz.ch

Abstract. We consider the problem of data association in a multi-person tracking context. In semi-crowded environments, people are still discernible as individually moving entities, that undergo many interactions with other people in their direct surrounding. Finding the correct association is therefore difficult, but higher-order social factors, such as group membership, are expected to ease the problem. However, estimating group membership is a chicken-and-egg problem: knowing pedestrian trajectories, it is rather easy to find out possible groupings in the data, but in crowded scenes, it is often difficult to estimate closely interacting trajectories without further knowledge about groups. To this end, we propose a third-order graphical model that is able to jointly estimate correct trajectories and group memberships over a short time window. A set of experiments on challenging data underline the importance of joint reasoning for data association in crowded scenarios.

Keywords: Grouping, Tracking, Data Association, Social Interaction.

1 Introduction

Tracking algorithms are an indispensable prerequisite for many higher-level computer vision tasks, ranging from surveillance to animation to automotive applications. Advances in observation models, such as object detectors or classification-based appearance models, have enabled tracking in previously infeasible scenarios. Still, tracking remains a challenging problem, especially in crowded environments. Tracking high numbers of pedestrians in such cases is even hard for humans. Usually, a manual annotator has to rely on higher-level reasoning, such as temporal information (that can go into the future) or social factors. Recent advances in the literature suggest that especially the latter can improve tracking performance. Typically employed social factors include a pedestrian's *destination*, *desired speed*, and *repulsion* from other individuals. Another factor is grouping behavior, which so far however has been largely ignored. For one, this is due to the fact that the grouping information (do two persons belong to the same group?) is not easily available. Still, groups constitute an

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part I, LNCS 6311, pp. 452–465, 2010.

important part of a pedestrian's motion. As we will show in this paper, people behave differently when walking in groups as opposed to alone: when alone, they tend to keep a certain distance from others, passing by closely only if necessary, but mostly at different speeds. When in groups, they try to stay close enough with other members, walking at the same speed.

In this paper, we therefore aim at exploiting the interaction between different people for data association in a principled way. In particular, we model group relations and study their effect on trajectory prediction. The grouping between pedestrians is treated as a latent variable, which is estimated jointly together with the trajectory information. Our model of choice is a third-order CRF, with nodes in the lower level corresponding to pedestrians, connected by third-order links that represent possible groupings. Recent advances in discrete optimization provide powerful tools for carrying out (approximate) inference in such models.

In order to take advantage of as much information as possible, we adopt a hypothesize-and-verify strategy over a frame-based tracking approach. By operating on short time windows (typically, in the order of a few seconds), useful statistics over pedestrians and group participation can be obtained, while only introducing a small lag as opposed to global trackers. The proposed framework operates in two steps: first, it generates possible trajectory hypotheses for each person within the given time window, then it selects the best hypothesis, taking into account social factors, while at the same time estimating group membership.

The paper is organized as follows: Related work is explored in Section 2. In Section 3, our model for the joint estimation of trajectories and groupings using social factors is introduced. The training of the model by natural statistics of interaction is discussed in Section 4. The inference method is then presented in Section 5. Finally, we show experimental results in Section 6, before concluding the paper in Section 7.

2 Related Work

Social behavior modeling. Based on a variety of psychological, physical, and social factors, people tend to keep certain distances from each other when interacting. Already investigated in the 1960ies by Hall [1] as proxemics, these factors are meanwhile also used for the modeling of pedestrian motion. Applications include simulation [2–4], computer graphics [5, 6], and, in the last few years, also Computer Vision [7–14]. While all of these works include various social factors, grouping information was mostly ignored. Helbing *et al*, in their seminal paper on the Social Force Model [2], include an attraction potential to model the group interaction. However, even in simulation applications, the notion of groups is rarely employed. In this paper, we will focus on group relations between subjects in a tracking setting, showing how to jointly estimate a person's correct trajectory and his group membership status. To the best of our knowledge, this paper is the first to explore the joint estimation of pedestrian trajectories and the grouping relations.

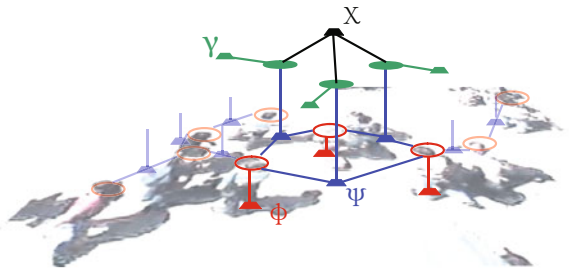

Fig. 1. Assumed higher-order model for joint trajectory and group finding (see text)

Tracking. Fostered by recent progress in object detection, there is an impressive body of work in single-person tracking-by-detection [15–20]. All propose different ways of handling the data association problem, but do not take advantage of any social factors beyond spatial exclusion principles. Only some researchers use social structures to improve tracking, most notably in crowded scenarios [7], or by modeling of collision avoidance-behavior [8, 11].

In this work, we focus on improving one building block of tracking—the data association—by taking advantage of social factors in a principled fashion. The proposed algorithm infers the best trajectory choice for each tracked object in a short time window. This thus means some latency as opposed to typical online trackers [15, 18, 19]. The method however does not need the entire time window either, as global approaches [17, 20]. [21] model simple interactions of targets in an MCMC framework, but not accounting for groups. [16] also use a hypothesize-and-verify strategy, however, they do not model any social factors, and a hypothesis contains an entire person's past, as opposed to a small window only. By operating in a shorter temporal window, our algorithm can take into account many more hypotheses, which is a requirement for tracking in challenging scenarios.

3 Group CRF

To improve data association in crowded scenarios, we want to jointly estimate pedestrian trajectories and their group relations. Fig. 1 shows the factor graph for the third-order CRF model we assume for this problem.

Given a starting frame, each tracking target i ($i = 1 \dots N$) is modeled as a variable node (red empty circle, Fig. 1), where each possible state corresponds to the choice of one local trajectory hypothesis $\mathbf{h}_i^m \in \mathcal{H}_i = \{\mathbf{h}_i^m\}^{m=1\dots M_i}$, with $\mathcal{H}_i$ the set of hypotheses for one person. As a trajectory hypothesis $\mathbf{h}_i^m$, we consider a single subject's possible future within a short time window. A joint assignment of hypotheses to all the subjects is defined as $\mathbf{H}^q = [\mathbf{h}_1^{q(1)} \dots \mathbf{h}_N^{q(N)}]$, where q is an assignment function that assigns each target i one hypothesis in $\mathcal{H}_i$ [1].

[1] To reduce notational clutter, we drop the superscripts for $\mathbf{h}$ and for $\mathbf{H}$ in the following.

The group variable $l_{c(ij)}$ (green filled circle, Fig. 1) indicates the group relation among the subjects i and j,

$$l_{c(ij)} = \begin{cases} 1 & \text{if subject } i \text{ and } j \text{ belong to the same group} \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

with $c(ij)$ an index function.

Two subjects i and j and the group variable $l_{c(ij)}$ are linked by a factor of order three (blue factor in Fig. 1). This link variable is essential to take advantage of grouping relations in our model. Grouping is an equivalence relation, i.e. it fulfills reflexivity, symmetry, and transitivity. While reflexivity and symmetry are enforced by the graph construction, the transitivity constraint is encoded in a third-order factor (black factor in Fig. 1): given three subjects i,j, and k for which there exist the link variables $l_{c(ij)}$, $l_{c(ik)}$, and $l_{c(kj)}$:

$$(l_{c(ij)} \wedge l_{c(ik)}) \rightarrow l_{c(kj)} \ . \tag{2}$$

The log-probability of a set of trajectories $\mathbf{H}$ and a set of grouping relations $\mathbf{L}$, given an image $\mathbf{I}$ and parameters Θ, is given by

$$\begin{aligned} \log P(\mathbf{H}, \mathbf{L}|\mathbf{I}, \Theta) = & \sum_i \phi_i^{motion}(\mathbf{h}_i|\Theta_{\phi^{motion}}) + \sum_i \phi_i^{app}(\mathbf{h}_i|\mathbf{I}, \Theta_{\phi^{app}}) + \\ & \sum_{c(ij)} \gamma_{c(ij)}(l_{c(ij)}|\Theta_\gamma) + \sum_{ijc(ij)} \psi^{pos}_{ijl_{c(ij)}}(\mathbf{h}_i, \mathbf{h}_j, l_{c(ij)}|\Theta_{\psi^{pos}}) + \\ & \sum_{ijc(ij)} \psi^{ang}_{ijl_{c(ij)}}(\mathbf{h}_i, \mathbf{h}_j, l_{c(ij)}|\Theta_{\psi^{ang}}) + \\ & \sum_{c(ij)c(ik)c(kj)} \chi_{c(ij)c(ik)c(kj)}(l_{c(ij)}, l_{c(ik)}, l_{c(kj)}|\Theta_\chi) - \log Z(I, \Theta), \end{aligned} \tag{3}$$

where ϕ^{app} and ϕ^{motion} model, respectively, the appearance and motion of a trajectory, $\gamma_{c(ij)}$ models the prior over a relation being of type group or not, $\psi^{pos}_{ijl_{c(ij)}}$, $\psi^{ang}_{ijl_{c(ij)}}$ model the grouping relation and $Z(I, \Theta)$ is the usual partition function making sure that the probability density function sums to one.

4 Learning the Parameters

Learning the parameters of the model in Eq. 3 could be done by maximizing the conditional likelihood of the parameters given the data. However, this is hard because of the partition function Z. Instead, inspired by piecewise training [22], we learn simple statistics from the data and define the terms in the Eq. 3 as a combination of these statistics. In particular we overparametrize the trajectory $\mathbf{h}_i$ as a sequence $[\mathbf{p}_i^0, s_i^0, \alpha_i^0 \ldots \mathbf{p}_i^{T-1}, s_i^{T-1}, \alpha_i^{T-1}]$ of , respectively, position, speed and orientation and extract simple statistics over these terms, rather than over the whole trajectory. In doing so, we use a non-parametric approach, by building histograms to estimate densities. The parameters Θ can be interpreted as the

Fig. 2. A snapshot from the sequence used in this paper. We only use data inside the red bounding box, to avoid stairs on the left and too heavy shadows in the upper part.

entries of these histograms. To reduce the notational clutter we will drop in the following the dependence on Θ.

In the analysis of the data, we make, when appropriate, a distinction between people walking and people standing. Besides believing that these two classes can have different statistics indeed, we are motivated for making this distinction by a technical limitation: the orientation estimate is hard and unreliable for standing people, while it can be approximated by the direction of motion for moving people. We therefore choose an empirical threshold of $0.15m/s$ [2] to distinguish between the two modes.

In the following, we will show the relevant statistics that we used in our model.

4.1 Dataset

The data used to extract the statistics has been kindly provided by Lerner *et al* [5]. The employed sequence shows a busy square from a stationary camera, oblique view, with a total of 450 subjects in 5400 frames. Most of the subjects walk from one of the borders of the scene to another and stay within the scene for about 15 seconds, while some stand longer in the scene talking to other subjects or waiting. An example frame is shown in Fig. 2. The sequence is particularly challenging due to low image resolution, interlacing and compression artifacts, cast shadows, as well as the large number of people. We manually annotated the head position of each subject and estimated a homography matrix to retrieve metric properties. In a second step, we annotated groups in the sequence, by relying on several cues, such as people talking to each other or holding hands, for example. For our purposes, we split the sequence in a training (3400 frames) and testing section (2000 frames).

4.2 Independent Motion and Appearance

Pedestrians change the walking direction smoothly. Furthermore, the walking speed is not arbitrary. This information is commonly exploited in motion prior

[2] Note that the estimation of the homography matrix introduces a small scalar factor compared to the real walking speed.

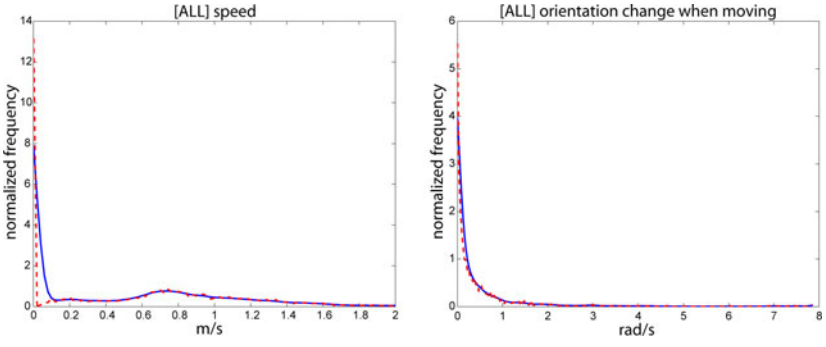

Fig. 3. Statistics over a person's movement: **Left:** the distribution $P(s_i^t)$ over speeds shows two peaks for people standing and walking. **Right:** the figure shows $P_{s_i^t \geq 0.15}(\alpha_i^t | \alpha_i^{t-1})$. For walking people, there is a preference to keep the current heading. Red indicates the original data points, blue the histogram estimate.

for pedestrians in a constant velocity model. To model these factors we define the motion term of Eq. 3 as

$$\phi_i^{motion}(\mathbf{h}_i) = \sum_{t=0}^{T-1} \log[P_{s_i^t < 0.15}(\alpha_i^t | \alpha_i^{t-1}) + P_{s_i^t \geq 0.15}(\alpha_i^t | \alpha_i^{t-1})] + \sum_{t=0}^{T-1} \log P(s_i^t) \ . \tag{4}$$

$P_{s_i^t < 0.15}(\alpha_i^t | \alpha_i^{t-1})$ is assumed uniform while $P(s_i^t)$ and $P_{s_i^t \geq 0.15}(\alpha_i^t | \alpha_i^{t-1})$ are estimated by building a normalized histogram (smoothed with a Gaussian kernel) of the angles and speeds extracted from the training set and are shown in Fig. 3. As one can expect, from the speed statistics it is easy to distinguish two modes, corresponding to standing and walking people. Fig. 3 shows also that the the choice of $0.15m/s$ for telling apart walking and standing pedestrian is a reasonable one.

For the appearance term, we directly use the output of the tracker (see Sec 6).

$$\phi_i^{app}(\mathbf{h}_i | \mathbf{I}) = \log f^{app}(\mathbf{h}_i | \mathbf{I}) \ . \tag{5}$$

4.3 Grouping Relations

Given two pedestrians, one of the obvious features that makes it possible to guess whether they belong to the same group or not, is proximity. So, when two pedestrians belong to the same group, their distance is kept to a certain value. If they are walking, the estimate of the orientation can give us further information on how they are positioned with respect to each other. For two pedestrians belonging to the same group, we therefore define

$$\psi_{ijl_{c(ij)}}^{pos}(\mathbf{h}_i, \mathbf{h}_j, l_{c(ij)} = 1) = \sum_{t=0}^{T-1} \log[P_{s_i^t \geq 0.15 \wedge s_j^t \geq 0.15}(\mathbf{p}_i^t | \mathbf{p}_j^t \alpha_j^t, l_{c(ij)} = 1,) + P_{s_i^t < 0.15 \vee s_j^t < 0.15}(d(\mathbf{p}_i^t, \mathbf{p}_j^t) | l_{c(ij)} = 1)] \ , \tag{6}$$

where $d(\mathbf{p}_i^t, \mathbf{p}_j^t)$ is the Euclidean distance between the positions $\mathbf{p}_i^t$ and $\mathbf{p}_j^t$. $P_{s_i^t \geq 0.15 \wedge s_j^t \geq 0.15}(\mathbf{p}_i^t | \mathbf{p}_j^t \alpha_j^t, l_{c(ij)} = 1)$ and $P_{s_i^t < 0.15 \vee s_j^t < 0.15}(d(\mathbf{p}_i^t, \mathbf{p}_j^t) | l_{c(ij)} = 1)$ are estimated using histograms as before and are shown in Fig. 4. For pedestrians do not belong to the same group, we found it unnecessary to distinguish between walking or standing. The main feature, when dealing with the position of two individual pedestrians, seems to be the *repulsion* effect: individuals try not to come close to each other unless necessary. In this case, we define the motion term as

$$\psi^{pos}_{ijl_{c(ij)}}(\mathbf{h}_i, \mathbf{h}_j, l_{c(ij)} = 0) = \sum_{t=0}^{T-1} \log P(d(\mathbf{p}_i^t, \mathbf{p}_j^t) | l_{c(ij)} = 0) \ , \tag{7}$$

where $P(d(\mathbf{p}_i^t, \mathbf{p}_j^t) | l_{c(ij)} = 0)$ is again estimated using histograms and shown in Fig. 4.

Another important feature of people when walking in the same group, is that they have the same orientation. We therefore define

$$\psi^{ang}_{ijl_{c(ij)}}(\mathbf{h}_i, \mathbf{h}_j | l_{c(ij)} = 1) = \sum_{t=0}^{T-1} \log P_{(s_i^t \geq 0.15 \wedge s_j^t \geq 0.15)}(\alpha_i^t, \alpha_j^t | l_{c(ij)} = 1) \ . \tag{8}$$

As before, this term is estimated with a smoothed histogram approach. The density is shown in Fig. 4 and, as expected, shows that subjects that walk together keep the same orientation. We did not observe an interesting orientation pattern among pedestrians that are not in the same group, therefore we assume uniform $P_{(s_i^t \geq 0.15 \wedge s_j^t \geq 0.15)}(\alpha_i^t, \alpha_j^t | l_{c(ij)} = 0)$.

Finally, $\gamma_{c(ij)}(l_{c(ij)})$ could be set by looking at the fraction of grouping relations over the total number of relations. Although the correct value for the fraction would be $\sim 23\%$ for our dataset, we will vary this value to measure the robustness of our model (see Sec. 6).

4.4 Transitivity Constraints

The hard constraint in Eq. 2 is modeled by penalizing impossible configurations with an opportunely large constant cost.

5 Inference

We are looking for the most probable joint assignment of the trajectories $\mathbf{H}$ together with the grouping relations $\mathbf{L}$ in Eq. 3. Exact inference is intractable, as the graph contains cycles and the potentials are not restricted to a particular kind (e.g., submodular). For the inference, we use Dual Decomposition (DD) [23], building on the code made available by [24]. DD optimizes the Lagrangian dual of the LP-relaxation of the original problem, by decomposing the problem into a set of subproblems, each of which can be solved efficiently. By optimizing the dual, it gives a lower bound that can be used to check whether the method converged to a global optimum (i.e. when the solution given by the primal has the same energy as the solution of the dual problem).

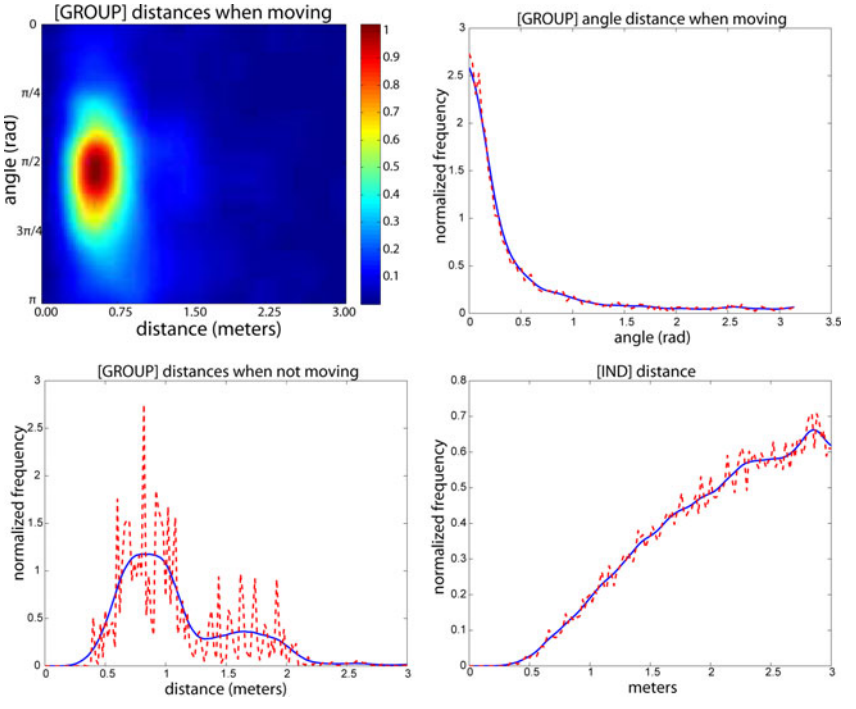

Fig. 4. Statistics over interacting people. **Top-left:** $P_{s_i^t \geq 0.15 \wedge s_j^t \geq 0.15}(\mathbf{p}_i^t | \mathbf{p}_j^t \alpha_j^t, l_{c(ij)} = 1)$ in polar coordinates, such that radius is the distance $d(\mathbf{p}_i^t, \mathbf{p}_j^t)$ and the angle is the angle under which j, with absolute orientation α_j^t sees i. When moving in groups, people keep a low distance from each other, trying to walk side by side. **Top-right** shows $P_{(s_i^t \geq 0.15 \wedge s_j^t \geq 0.15)}(\alpha_i^t, \alpha_j^t | l_{c(ij)} = 1)$. As expected, people that walk together are headed in the same direction. **Bottom-Left** shows $P_{s_i^t < 0.15 \vee s_j^t < 0.15}(d(\mathbf{p}_i^t, \mathbf{p}_j^t) | l_{c(ij)} = 1)$. The distribution is less peaked than the distribution shown in the top-left figure, probably reflecting the fact that when people are standing in groups, they allow for more flexible *configurations*. **Bottom-Right**: the figure shows $P(d(\mathbf{p}_i^t, \mathbf{p}_j^t) | l_{c(ij)} = 0)$. Like for groups, the repulsion effect between individuals, used in many pedestrian motion models [2, 11], is evident from the low value around 0.

In our case, we decompose the original graph first into a constraint layer containing only transitivity constraints factors and a data layer containing all the other factors. Then these sub-graphs are further decomposed into spanning trees. We optimize each tree separately using standard Belief Propagation [25]. The primal solution, and therefore the upper bound to the optimal solution, is found by using a heuristic similar to that described in [23].

6 Experiments

The proposed model requires a set of hypotheses to choose from. In this section, we therefore first describe how to build up the model given an input frame, before presenting experiments on real-world data.

6.1 Model Construction

Hypothesis Generation. Given a starting frame t_0, a separate set of hypotheses $\mathcal{H}_i$ is generated for each currently tracked person i. Each hypothesis $\mathbf{h}_i$ describes a possible motion of the subject between $t = t_0 \dots t_{T-1}$. To this end, we start a single-person tracker for each person i at t_0, at each time step following the cost function recursively according to a best-first paradigm. Following at each time step t the M best options therefore yield a maximum of M^T hypotheses per person. As a cost function, we employ several cues: as a motion and appearance model, we use a constant velocity assumption, respectively an HSV-color histogram $\mathbf{a}_i^t$ on the subject's head. The product of the Bhattacharyya coefficients $d(\cdot,\cdot)$ along the trajectory is then used to define $f^{app}(\mathbf{h}_i|\mathbf{I}) = \prod_{t=1}^{T-1} d(\mathbf{a}_i^t|\mathbf{a}_i^{t-1})$.

As a third cue, we consider a discrete set of detections in the current hypothesis' vicinity. The detections are obtained from a voting-based detector [26], trained on both head and upper bodies from a total of 1145 positive and 1208 negative examples. Even though specifically trained on the same setup's data, the detector only reaches an equal error rate of 0.65 (head) respectively 0.76 (torso) (see Fig. 5). The reason for this low performance is a higher number of false positives on strong cast shadows, as well as some false negatives when people are standing very closely together. To account for frequent false negatives, up to 50% of missing detections are allowed inside a trajectory, where the missing parts are interpolated using the constant velocity model.

To handle the case of persons leaving the scene, we introduce a set of virtual detections at the border of the image. Once a tracker selects such a detection, it is terminated, and the corresponding trajectory corresponds to a linear extrapolation starting from that time step.

For computational reasons, in the presented experiments, we use a time step of 0.2 seconds, and set $T = 10$ (thus always considering time windows of 2 seconds) and $M = 4$, yielding an average of 147 hypotheses per subject. We run

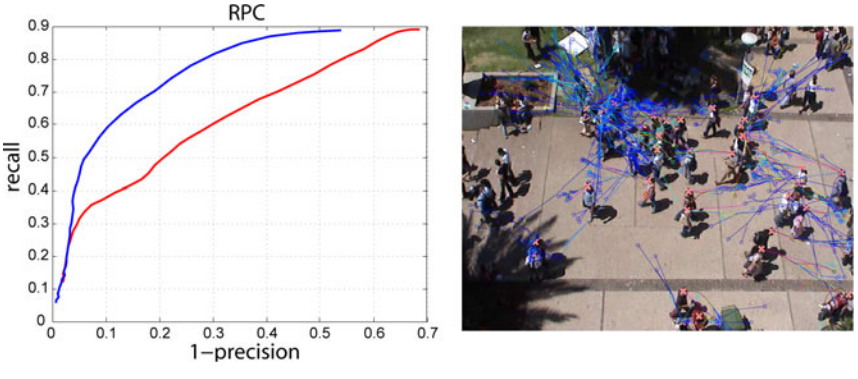

Fig. 5. Left: RPC curves for head detector (red) and torso detector (blue). **Right:** Sample hypotheses for one frame, with blue corresponding to low confidence and red to high. Especially in crowded areas, many possible hypotheses can be generated.

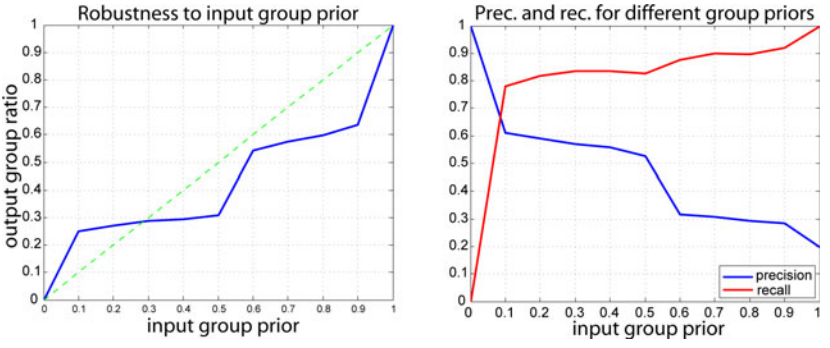

Fig. 6. Left: Estimate of groups relations while varying the grouping prior γ. **Right:** precision and recall curves for group relations for different γ values.

the experiment each 2 s for all pedestrians, starting from 40 different frames. This results in 1236 subjects being tracked.

Link Modeling. To set up the links between individuals, Delaunay triangulation is performed on the subjects positions in the input frame. Links longer than 3 meters are cancelled.

6.2 Groundtruth

Before using an actual detector to drive hypothesis generation, we perform a baseline experiment, where we use the ground-truth annotations as detections (note that we are operating on the test sequence, i.e., the training of the model did not use this data at all). To measure the effect of the proposed model, we compare the output of the inference stage with locally selecting the best trajectories (i.e., the hypothesis with the maximum unary term). We report the number of correctly selected trajectories as the ones that coincide with the ground-truth completely. In Fig. 6 (left), we run this experiment for different values of the grouping prior γ. As can be seen, the model does not blindly trust the prior (unless set to the extreme positions), but moves towards the true fraction of groupings (0.23) disregarding the starting position. The performance of the model with respect to trajectory selection is hardly affected by the grouping: the unary makes 34 mistakes, whereas the full model, depending on the chosen grouping prior, performs considerably better with 12 ± 2 mistakes. Only when $\gamma = 0$, our model makes 22 mistakes. In Fig. 6 (right), we furthermore plot recall and precision of finding groups, again varying over the prior γ. The numbers stay quite constant for a large range of γ, underlining the stability of the model. Choosing extreme values will naturally also lead to inferior results, either in favor of groups or not. In the upcoming experiments, we will use an uninformed prior, $\gamma = 0.5$.

Table 1. Performance of model when using raw detections as input. The proposed model not only improves the correctly chosen trajectories, but also recovers groups with high recall and good precision.

	Wrong Trajectories	Groups			
		TP	TN	FP	FN
Local	401	-	-	-	-
Group CRF	363	389	1526	449	84

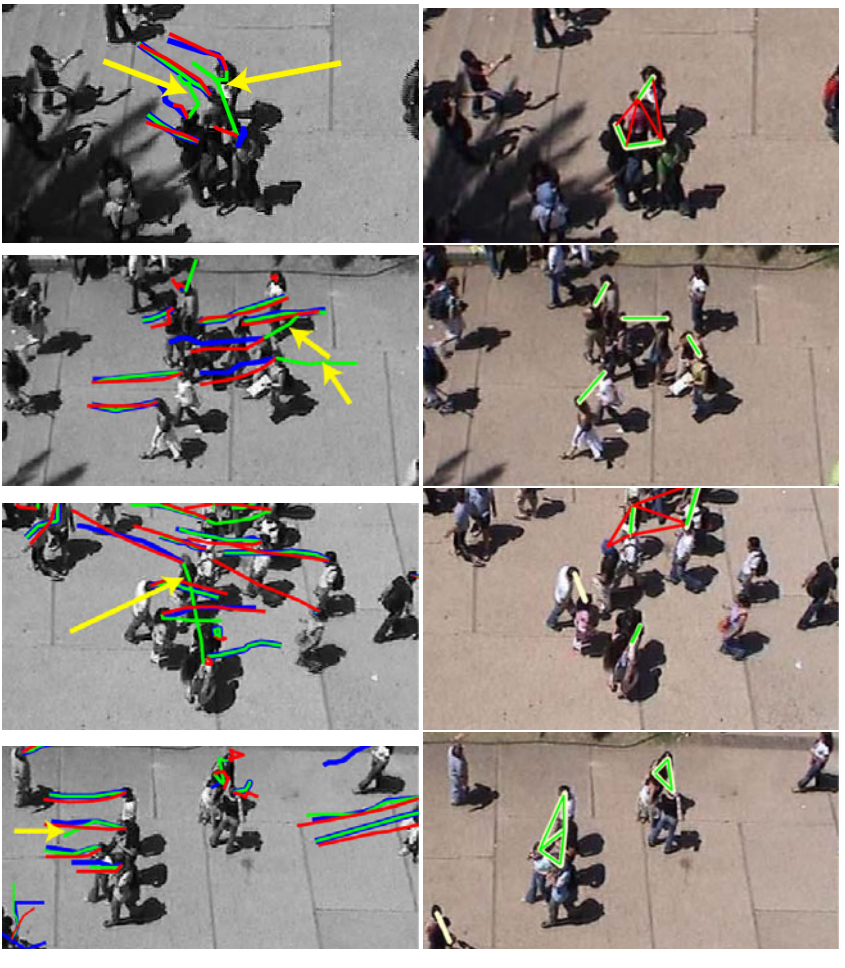

Fig. 7. Example situations (close-ups). **Left:** trajectories, with ground truth (red) and solutions found by the unary term alone (green) and the group CRF (blue). **Right:** grouping, with ground truth (white), true/false positives (green/red) (see text).

6.3 Detector

When starting from a ground-truth point and generating hypotheses using detections, the generation step has to deal with a considerable number of false positives (generating excess wrong trajectories) and false negatives (in the worst case, missing an entire trajectory). Due to these inaccuracies, we change the notion of correct trajectory to an error $< 0.5\,\mathrm{m}$ from the ground-truth at the last trajectory position. The subject errors and the group statistics are reported in Table 1. Note that this experiment is considerably harder, so the number of errors in absolute terms increases. Still, our method improves $\approx 10\%$ w.r.t. using only the unary terms, i.e. without grouping. The group statistics show a precision of 46% (about twice above the chance level of 23%) and 82% recall.

Some example images, comparing the two methods, are shown in Fig. 7. For each sample, we report both the trajectories found by either choosing the local optimum or the group CRF, as well as the recovered grouping by our model. In the top row, the grouping information gives a twofold improvement, encouraging the two persons to move together to the left side, as opposed to choosing intersecting trajectories (yellow arrows). One single wrong link between the two correctly inferred groups spurs the creation of additional wrong links through transitivity. In the second row, grouping correctly enforces the two people in the middle to walk together to the left as opposed to the local solution, which erroneously goes to the right (yellow arrows). In the third row, the joint reasoning keeps the group CRF from choosing the wrong path leading through all the pedestrians (yellow arrows), thus highlighting the spatial exclusion constraint. Finally, in the last row, grouping encourages smoother trajectories that stay well separated, with the group on the left correctly estimated.

7 Conclusions

In this paper we investigated the influence of pedestrian interactions on data association in crowded scenes, having in mind a tracking application. Statistics learned on natural video data show that people walking in groups behave differently from people walking alone. Commonly hard-coded effects such as repulsion/avoidance were also clearly visible in the data. These statistics were used to train a graphical model encoding the interactions between pedestrians in a principled manner. The model was optimized for the MAP estimate with a state of the art approximate inference engine, giving a joint estimate about correct trajectories and group memberships in the data.

The results show that interactions should be taken into account when reasoning about people trajectories. We not only showed that joint optimization is beneficial in terms of tracking error, but we were able to recover, with a good recall and a sufficient precision, group statistics.

The running time depends on the number of people in the scene. Our current implementation of the system, far from being optimized, takes few minutes (≈ 10) to output trajectories of length 2 seconds and grouping relations.

The focus of this paper was rather the effect of interactions as opposed to a complete tracking application. We therefore only showed results on short time windows initialized from ground-truth locations, not forming entire trajectories automatically. Extending the model in this direction will be part of future work.

References

1. Hall, E.T.: The Hidden Dimension. Garden City (1966)
2. Helbing, D., Molnár, P.: Social force model for pedestrian dynamics. Physical Review 51(5) (1995)
3. Penn, A., Turner, A.: Space syntax based agent simulation. In: Pedestrian and Evacuation dynamics (2002)
4. Schadschneider, A.: Cellular automaton approach to pedestrian dynamics - theory. In: Pedestrian and Evacuation Dynamics (2001)
5. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. In: EUROGRAPHICS (2007)
6. Massive Software: Massive (2010)
7. Ali, S., Shah, M.: Floor fields for tracking in high density crowd scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 1–14. Springer, Heidelberg (2008)
8. Antonini, G., Martinez, S.V., Bierlaire, M., Thiran, J.: Behavioral priors for detection and tracking of pedestrians in video sequences. IJCV 69, 159–180 (2006)
9. Choi, W., Shahid, K., Savarese, S.: What are they doing? collective activity classification using spatio-temporal relationship among people. In: Workshop on Visual Surveillance (VSWS '09) in conjuction with ICCV'09 (2009)
10. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using Social Force model. In: CVPR (2009)
11. Pellegrini, S., Ess, A., Schindler, K., Gool, L.V.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: ICCV (2009)
12. Scovanner, P., Tappen, M.: Learning pedestrian dynamics from the real world. In: ICCV (2009)
13. Ge, W., Collins, R., Ruback, B.: Automatically detecting the small group structure of a crowd. In: IEEE Workshop on Applications of Computer Vision, WACV (2009)
14. French, A.: Visual Tracking: From An Individual To Groups of Animals. PhD thesis (2006)
15. Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., Gool, L.V.: Robust tracking-by-detection using a detector confidence particle filter. In: ICCV (2009)
16. Ess, A., Leibe, B., Schindler, K., Gool, L.V.: A mobile vision system for robust multi-person tracking. In: CVPR (2008)
17. Li, Y., Huang, C., Nevatia, R.: Learning to associate: HybridBoosted multi-target tracker for crowded scene. In: CVPR (2009)
18. Okuma, K., Taleghani, A., de Freitas, N., Little, J., Lowe, D.: A boosted particle filter: Multitarget detection and tracking. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 28–39. Springer, Heidelberg (2004)
19. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet part detectors. IJCV 75, 247–266 (2007)
20. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: CVPR (2008)

21. Khan, Z., Balch, T., Dellaert, F.: MCMC-based particle filtering for tracking a variable number of interacting targets. PAMI 27(11), 1805–1819 (2005)
22. Sutton, C., McCallum, A.: Piecewise training of undirected models. In: Conference on Uncertainty in Artificial Intelligence, UAI (2005)
23. Komodakis, N., Paragios, N., Tziritas, G.: MRF optimization via dual decomposition: Message-passing revisited. In: ICCV (2007)
24. Torresani, L., Kolmogorov, V., Rother, C.: Feature correspondence via graph matching: Models and global optimization. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 596–609. Springer, Heidelberg (2008)
25. Mooij, J.M., et al.: libDAI 0.2.5: A free/open source C++ library for Discrete Approximate Inference (2010), `http://www.libdai.org/`
26. Gall, J., Lempitsky, V.: Class-specic hough forests for object detection. In: CVPR (2009)