Socially-aware Large-scale Crowd Forecasting

Alexandre Alahi* Vignesh Ramanathan† Li Fei-Fei*
*Computer Science Department, Stanford University
†Department of Electrical Engineering, Stanford University
{alahi, vigneshr, feifeili}@cs.stanford.edu

Abstract

In crowded spaces such as city centers or train stations, human mobility looks complex, but is often influenced only by a few causes. We propose to quantitatively study crowded environments by introducing a dataset of 42 million trajectories collected in train stations. Given this dataset, we address the problem of forecasting pedestrians' destinations, a central problem in understanding large-scale crowd mobility. We need to overcome the challenges posed by a limited number of observations (e.g. sparse cameras), and change in pedestrian appearance cues across different cameras. In addition, we often have restrictions in the way pedestrians can move in a scene, encoded as priors over origin and destination (OD) preferences. We propose a new descriptor coined as Social Affinity Maps (SAM) to link broken or unobserved trajectories of individuals in the crowd, while using the OD-prior in our framework. Our experiments show improvement in performance through the use of SAM features and OD prior. To the best of our knowledge, our work is one of the first studies that provides encouraging results towards a better understanding of crowd behavior at the scale of million pedestrians.

1. Introduction

Recent studies have shown that our mobility is highly predictable at a city-scale level [31]. The location of a person at any given time can be predicted with an average accuracy of 93% supposing 3 km^2 of uncertainty. How about at a finer resolution such as in shopping malls, in airports, or within train terminals for safety or resource optimization? What are the relevant cues to best predict human behavior? Kitani *et al*. [16] have shown that scene semantics is a strong cue to forecast pedestrian's trajectory. Previous work [12, 22] has also shown that our mobility is influenced by our neighbors, either consciously, *e.g*. by relatives or friends, or even unconsciously, *e.g*. by following an individual to facilitate navigation. In public spaces, both low and high density crowds are observed leading to the following challenges to capture and forecast human mobility: (i) people highly occlude each other making appearance cues less discriminative, (ii) the independent motion prior [15, 6, 1] becomes a weak assumption in crowds since social interactions can influence human dynamics, and (iii) observations are often limited since a sparse and scattered network of cameras is usually installed. In this paper, we address the above challenges by proposing a forecasting algorithm leveraging fine and coarse priors to predict crowd behavior. We propose a new descriptor, called as Social Affinity Map (SAM), to address the lack of appearance information and the weak independent motion prior in linking tracklets[1] from a sparse network of cameras.

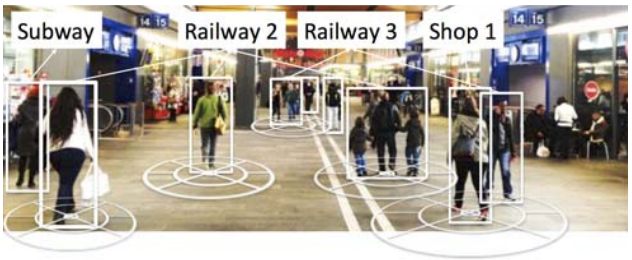


Figure 1: Predicting the behavior of pedestrians given Social Affinity Maps (SAM). We define SAM as a radial binary descriptor representing the spatial configuration of your neighbors.

There are three levels of understanding mobility according to [13]: strategic level (intended goal), tactical level (route choice), and operational level (actual movement at each time instant). We propose to study the latter operational level to forecast the former strategic one. In other words, we model the social interactions of pedestrians to predict their destination (see Fig. 1). Forecasting destinations is often referred to as estimating Origin-Destination (OD) Matrices [20]. It represents the starting and ending points of all pedestrian trajectories during a time period.

The key contributions of our paper are as follows:

1. We introduce a large-scale dataset of 42 million trajectories extracted from real-world train stations.

[1] A tracklet is a track fragment captured by a single camera with high confidence.

1063-6919/14 $31.00 © 2014 IEEE
DOI 10.1109/CVPR.2014.283

IEEE computer society

2. We propose a new feature descriptor to capture the behavioral signature of neighbouring pedestrians termed as Social Affinity Map (SAM).

3. We formulate the problem as a linear integer program using the proposed SAM feature along with an OD prior, and we present a heuristic optimization method to solve it.

1.1. Large-scale data collection campaign

Social interactions are relationships between individuals, which might not occur in all environments. In order to model and best understand them, we need to capture real-world rather than simulated data at large-scale. Therefore, we have installed a dense network of more than a hundred cameras in train stations to capture the full trajectories of pedestrians (see Fig. 2). At any given time, up to a thousand pedestrians can be within the same area (*e.g*. the corridor illustrated in Fig. 2). Such a data collection campaign allows us to validate the occurrence of social affinities and their impact on forecasting real-world pedestrians' behavior over 42 million trajectories. We share the captured dataset[2] to enable various research communities, from psychology to computer vision, to dive into a large-scale analysis of human mobility in crowded environments.

Collecting the behavior of pedestrians in a network of cameras implies the following steps: (i) Detection, (ii) Tracklet generation, (iii) Tracklet association. For (i) and (ii), we use state-of-the-art detection [11] and tracklet generation algorithms [19]. Briefly, to achieve high confidence on the detection performance, we have installed top-view optical and thermal imaging to be robust to illumination changes and prevent self occlusions. When a top-view was not possible (due to low ceiling), we installed depth cameras (rgb-d sensors) to capture 3D detection given partial occlusions. We evaluated the detection performance over ten thousands manually labeled pedestrians. This led to 95% recall with more than 99% precision, thanks to the controlled viewpoints and sensing modalities. We use a sparsity driven framework to segment foreground silhouettes, given a dictionary of pre-computed ideal silhouettes [1]. Once people are located on the ground, we solve the minimum network flow presented by Leal-Taixe *et al*. [19]. The last (iii) tracklet association step will be discussed in Section 4 network of cameras.

2. Related work

Large-scale pedestrian tracking. Past decades have witnessed many datasets dealing with camera networks for various applications such as sports analysis [7], security (PETS workshops) [10], traffic modeling [29] and more recently the video understanding dataset [8], to name a few. They consist of several hours of video sequences. However, they do not address large-scale setups, where the Origin-Destination of long-term tracks are of interest in crowded environments. Crowd behavior has usually been addressed only with a single camera monitoring part of a marathon, or a political rally, where OD analysis is limited [2, 28].

Tracklet association. A large body of work models visual appearance to link tracklets across cameras [17, 9, 34, 26]. Andriluka *et al*. [3] use person detection as a cue to perform tracking and vice-versa. Javed *et al*. in [14] use travel time and the similarity of appearance features. Song *et al*. in [30] use a stochastic graph evolution strategy. Tracklets extracted by each camera are linked with the Hungarian algorithm [25], MCMC [33], or globally optimal greedy approaches [26]. These approaches have not addressed the linking of tracklets that are dozens of meters away in a highly crowded scene. In addition, given the camera viewing angle and the pedestrian flow, often only part (*e.g*. the hair) of a pedestrian's body is captured. Therefore, other cues need to be exploited. However, these methods do not capture the social interactions, which are valuable cues in crowded and long-distance settings.

Tracking with social prior. Social behavior has recently been incorporated into existing tracking frameworks by modeling the well-known social forces [12] with Kalman filters [21], extended Kalman filters [23], or Linear Programming [19, 27]. Antonini *et al*. [4] use Discrete Choice Models to simulate the walking behavior of people. These approaches improve the operational-level tracking when a few frames are missing (*e.g*. when given a low-frame rate, or short occlusion cases). They also often model a grouping cue to solve the data association problem [24, 19, 27]. They model it as a set of pedestrians with similar velocities and spatial proximity. Similarly, [18] use grouping cues in a hierarchical framework to identify sports player roles. The grouping cue is typically handled as a binary variable indicating group similarity. However, the key challenge is to use a finer representation to capture group association and integrate it into the problem of tracklet association. Yang *et al*. [32, 24] use a conditional random field framework to jointly estimate group membership and tracks. Leal *et al*. [19] iteratively compute the minimum cost flow for various velocity and grouping assignments until convergence or when a maximum number of iterations is reached. Qin *et al*. [27] use the Hungarian algorithm to jointly group and link tracklets. However, the Hungarian algorithm does not solve the global minimization over the full long-term track, whereas the minimum network flow formulation does. In this work, we propose a descriptor representing the grouping cue as a feature to efficiently match behavior across pedestrians.

3. Social Affinity Map: SAM

Our collected dataset enables us to study human behavior in crowded settings. In this paper, we focus our analysis

[2] www.ivpe.com/crowddata.htm

Monitoring area	Density	Travel time	Distance	Speed	Traffic / day	Total
$20.000m^2$	up to 1 ped/m^2	1 min	100 m	1.37 m/s	100.000-250.000/terminal	42 million trajectories

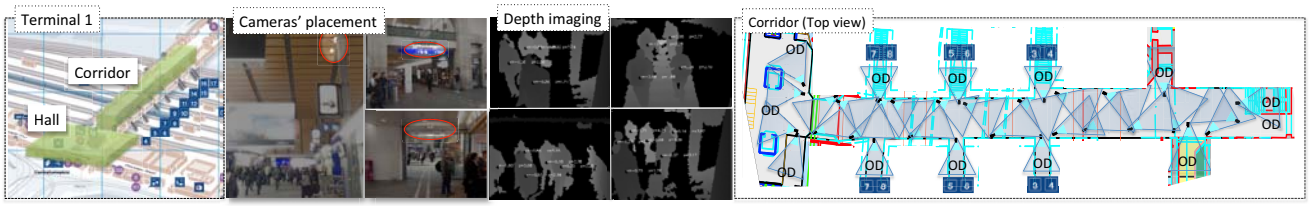


Figure 2: Real-world setup. Top row presents some facts regarding the dataset (values are in average). Bottom row illustrates one of the monitored corridors. More than 30 cameras are deployed in the presented corridor, whereas 132 cameras are deployed in total in 3 corridors, one track, and one large hall. At any given time, the occupancy of the corridor can reach more than one thousand of pedestrians. The label "OD" represents entry/exit zones.

on *social affinity* [3], which bonds people together in large crowds. We are interested in behavioral cues that remain stable over time and across various sensing modalities (e.g. optical, thermal, and depth) to link far-away tracklets.

Definition 1 *We define "social affinity" as the motion affinity of neighboring individuals.*

Social affinities can be consciously formed by friends, relatives or co-workers. However, in crowded environments, subconscious affinities exist. For example, the "*Leader-follower*" phenomenon [22] represents a spontaneous formation of lanes in dense flows, as a result of fast pedestrians, passing slower ones. More formally, the leader-follower pattern captures the behavior of a pedestrian (a follower) who adjusts his/her motion to follow a leader to enable smooth travel. We propose to learn the various social affinities which bind people in a crowded scene through a new feature called as Social Affinity Map (SAM).

3.1. The SAM feature

We observed that in public settings, social forces are mostly determined by the proximity of people to each other as noted in previous works [12]. Since, people are more easily influenced by others in their vicinity, we develop a social affinity feature which captures the spatial position of the tracklet's neighbors. As shown in Fig. 3, we achieve this by radially binning the position of neighbouring tracklets.

We further learn the spatial binning by first clustering the relative position of surrounding individuals over all captured trajectories. We considered relative positions within a limit of $3m$, to avoid outliers. The distribution of the relative positions across the million trajectories is visualized in Fig. 3. We obtain 10 bins as a result of this clustering, as shown in the figure. The percentage of relative positions pooled into this bins is also shown in the figure. It is interesting to point out that the most used bin is the one on far right side ("N" label in Fig. 3). It can be interpreted as the comfortable pattern to walk with respect to other individuals as opposed to the left hand side.

[3] Additional analysis can be found in the supplementary material.

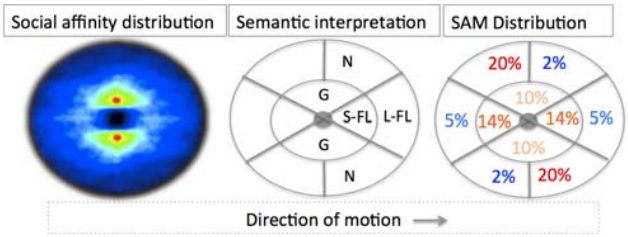


Figure 3: Left hand-side: Heatmap of the relative positions of all neighboring pedestrians across all tracklets. Middle column: it represents the SAM with our semantic description where "G" is the group affinity (such as couples, friends), "S-FL" is the short distance Follow-Leader behavior, "L-FL" is the long distance FL behavior, and "N" can be seen as the comfortable distance to maintain while walking in the same direction. The right hand-side represents the distribution of presented behavior.

Given a new tracklet, we perform vector quantization (VQ) coding to obtain the SAM feature. We fit a Gaussian Mixture Model to the relative position of its surrounding tracklets. The inferred GMM values within the previously learned spatial bins are discretized to obtain a binary radial histogram, which represents the SAM feature vector. The complete process is illustrated in Fig. 4. Hamming distance is used to compare SAM across tracklets. Note that binary quantization has little impact on the efficacy of the feature, and is only used to speed up the comparison method.

Our SAM feature can differentiate between various configurations of social affinities such as "*couple walking*", or the "*Leader-follower*" behavior. Fig. 5 illustrates the 8 most observed SAM over millions of trajectories. It is worth pointing out that 76% of individuals belong to a group, hence a SAM provides valuable information in crowded settings, motivating the use of these cues in forecasting the mobility of pedestrians.

4. Forecasting mobility: problem formulation

We have a sparse network of cameras monitoring the transit of people in a public setting like a railway terminal. The terminal has a set of entry points referred to as the *ori-*

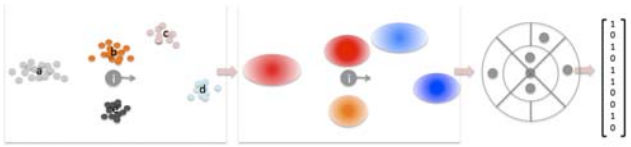


Figure 4: Illustration of a Social Affinity Map extraction (top view). The relative positions of neighboring individuals are clustered into a radial histogram. The latter is one bit quantized.

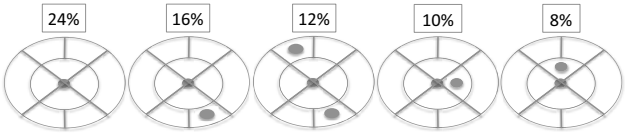


Figure 5: Illustration of the 8 most observed social affinities learned from the data. The above percentage represents the frequency of occurrence of the corresponding SAM.

gin, and exit points referred to as the *destination*. The goal of our work is to identify the Origin and Destination (OD) of every person entering and exiting the camera network. We achieve this by identifying the trajectories which connect the tracklets starting at the origin to the tracklets ending at the destination. The number of intermediate tracklets linked to obtain these trajectories decreases with the sparsity of the camera network. Fig. 6 illustrates an extreme case with only origin and destination tracklets.

We have a set of origin tracklets O and an equal number of destination tracklets D. Each tracklet in O is captured at one of the many entrances into the area, and a destination track in D is captured at an exit. We also have a set of intermediate tracklets X obtained by our sparse camera network. We want to find the set of trajectories T, where each trajectory $t \in T$ is represented as an ordered set of tracklets, (o_t, X_t, d_t), with $o_t \in O$ and $d_t \in D$ representing the origin and destination tracklets of the trajectory. Similarly, $X_t = (x_t^{(1)}, \ldots, x_t^{(n)})$ is an ordered set of intermediate tracklets which are linked to form the trajectory. These tracklets are ordered by the time of initiation. The problem can be written as a Maximum a-posteriori estimation problem similar to [34, 26]:

$$T^* = \max_T P(X|T)P(T), \tag{1}$$

where $P(X|T)$ is the probability of the tracklets in X being true positive tracklets. The probability $P(X|T)$ is:

$$P(X|T) \propto \prod_{t \in T} \prod_{x \in X_t} \frac{P_{tp}(x)}{P_{fp}(x)}, \tag{2}$$

where $P_{tp}(x)$ and $P_{fp}(x)$ are probabilities of the tracklet being a true positive, and false positive respectively.

We define $P_{OD}(o, d)$ as the OD-prior term which states the probability of a person entering at the origin corresponding to o exiting at the destination corresponding to d. Such prior is often neglected and assumed to be uniform. However, in many applications, it is a strong prior, such as avoiding forbidden paths in airports.

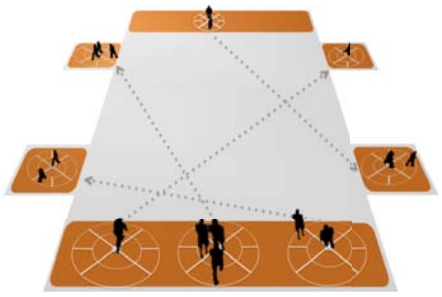


Figure 6: Predicting the behavior of pedestrians given Social Affinity Maps (SAM) with few cameras. Orange regions represent the monitoring areas of cameras. We illustrate the extreme case when cameras are only placed at entrance or exit zones, referred to as OD cameras.

Next, similar to [26], we assume a Markov-chain model connecting every intermediate track $x_t^{(i)}$ in the trajectory T, to the subsequent track $x_t^{(i+1)}$ with a probability given by $P(x_t^{(i+1)}|x_t^{(i)})$. The trajectory probability $P(T)$ is:

$$\begin{aligned} P(T) &= \prod_{t \in T} P(t), \\ P(t) &= P_{OD}(o_t, d_t) P\left(x_t^{(1)}|o_t\right) \\ &\quad \prod_{i=2}^{n} P\left(x_t^{(i)}|x_t^{(i-1)}\right) P\left(d_t|x_t^{(n)}\right), \end{aligned} \tag{3}$$

where $n = |X_t|$ is the number of intermediate tracklets in the trajectory.

The MAP problem from Eq. 1 can now be formulated as a linear integer program in a manner similar to [26]:

$$\begin{aligned} \min_f \quad & C(f) \\ C(f) \quad = \quad & \sum_{x_i \in X} \alpha_i f_i + \sum_{x_i, x_j \in X} \beta_{ij} f_{ij} + \\ & \sum_{\substack{x_i \in X, \\ o \in O}} \beta_{oi} f_{oi} + \sum_{\substack{x_i \in X, \\ d \in D}} \beta_{id} f_{id} + \sum_{\substack{o \in O, \\ d \in D}} \gamma_{od} f_{od} \\ \text{s.t} \quad & f_i, f_{ij}, f_{od} \in \{0, 1\} \\ \text{and} \quad & f_i = \sum_j f_{ij} + \sum_d f_{id} = \sum_i f_{ji} + \sum_o f_{oi}, \\ & \sum_{od} f_{od} = |O| = |D|, \\ & \sum_d f_{od} = \sum_i f_{oi}, \\ & \sum_o f_{od} = \sum_i f_{id} \;\; \forall \, x_i, x_j \in X, \; o \in O, \; d \in D, \end{aligned} \tag{4}$$

where f_i is the flow variable indicating whether the corresponding tracklet is a true positive, and f_{ij} indicates if the corresponding tracklets are linked together. The variable β_{ij} denotes the transition cost given by $\log P(x_i|x_j)$ for the tracks $x_i, x_j \in X$. The log-likelihoods β_{oi}, β_{id} are also defined similarly, for the origin track o and destination track d. The local cost α_i is the log-likelihood of an intermediate track being a true positive. Finally, the OD-prior cost is represented as $\gamma od = \log P_{OD}(o, d)$.

We note that the optimization problem in Eq. 4 is equivalent to the flow optimization problem widely discussed in [26, 34] in the absence of the OD prior term. Such problems can be solved through k-shortest paths or the more efficient greedy approach proposed in [26]. However, the addition of the OD-prior term leads to loops in the network-flow problem, and can no longer be solved exactly through shortest path algorithms. Hence, we adopt a heuristic approach to solve Eq. 4, as discussed in Sec. 5.

4.1. Local cost

The local cost α_i is proportional to the length of a tracklet. This helps us to remove short tracklets that might represent false positives.

4.2. Transition cost

The transition cost β_{ij} for any two tracklets is split into two components as shown below.

$$\beta_{ij} = \beta_{ij}^{SAM} + \beta_{ij}^{M}, \tag{5}$$

where β^{SAM} is the social-affinity cost and β^{M} is a cost to ensure smoothness in the connected tracklets.

Social Affinity cost. In our model, we wish to ensure that tracklets moving in similar social groups have a stronger likelihood of being linked to each other. This affinity forms an important component in large scale tracking scenarios like ours, where the appearance of an individual is not very discriminative. The SAM features introduced in Sec. 3 are used to measure the social affinity distance between tracklets moving in groups as shown below

$$\beta_{ij}^{SAM} = \mathbf{H}(sam_i, sam_j), \tag{6}$$

where $\mathbf{H}(.)$ denotes the Hamming distance between two binary vectors, and sam_i, sam_j denote the SAM feature vector of the two tracks.

Motion similarity. Another cue β^{M}, which is used to ensure smoothness in trajectory motion is obtained by measuring the distance between the motion patterns of two tracklets similar to [33, 30]

4.3. OD-prior cost

The OD-prior cost is the log-likelihood of the prior probability of transiting from an origin point to the destination. In most surveillance settings, we can use prior knowledge on the geography of the terminal, as well as rough estimates of the passenger freight to obtain an OD prior. In addition, the OD prior can be used to enforce constraints such that passengers entering a certain entry point would not return to the same location from a parallel entrance. In our experiments in later sections, the OD prior is obtained by a short survey in the location. This prior will also be released along with the dataset.

5. Optimization

As stated before, the optimization in Eq. 4 cannot be trivially solved through existing shortest path algorithms [26] as in the case of traditional tracking. Hence, we adopt a heuristic approach as explained below.

Greedy optimization with OD-prior. We first run a greedy algorithm to identify the low-cost solutions in the graph:

1. Find the shortest path which links an origin tracklet to the destination tracklet in Eq. 4
2. Remove the tracklets which are part of the trajectory obtained in the previous step and repeat.

The greedy algorithm provides an approximate solution to the problem and is computationally efficient. However, it does not solve the global optimization problem. We use a simple heuristic explained below to obtain a better solution.

Optimization with OD re-weighted cost. The solution of the greedy algorithm helps us identify the paths which agree with the OD-prior. Hence, the transition flow variables set by this algorithm provide a rough estimate of the pairwise affinity between tracklets in the presence of OD-prior. We use this intuition to add an additional cost which penalizes the link between tracklets which were not originally connected by the greedy algorithm. While adding this cost, we remove the original OD-prior cost γ_{od}, thus resulting in a network-flow problem which can be solved by k-shortest path approach. The modified cost $\tilde{C}$ is shown below:

$$\begin{aligned}\tilde{C}(f) \quad = \quad & \sum_{x_i \in X} \alpha_i f_i + \sum_{x_i, x_j \in X} \tilde{\beta}_{ij} f_{ij} + \\ & \sum_{\substack{x_i \in X, \\ o \in O}} \tilde{\beta}_{oi} f_{oi} + \sum_{\substack{x_i \in X, \\ d \in D}} \tilde{\beta}_{id} f_{id},\end{aligned} \tag{7}$$

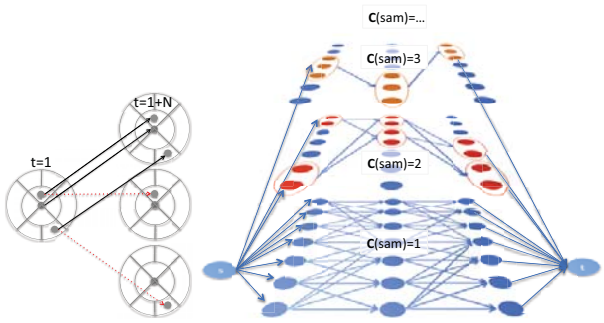
where $\tilde{\beta}$ is the OD-re-weighted cost defined below.

$$\tilde{\beta}_{ij} = \beta_{ij} + \lambda \mathbf{1}(f_{ij}^{greedy} = 1), \tag{8}$$

where f_{ij}^{greedy} is the solution obtained from the greedy algorithm and λ is a parameter indicating the strength of the OD-prior cost. The transition cost is re-weighted for all pairs of tracklets including the origin and destination tracklets.

6. Coarse-to-Fine Data Association

The model presented in Sec. 4, uses a social affinity cost to ensure that tracklets with similar grouping cues are con-



(a) Toy example

(b) Coarse-to-fine data association

Figure 7: (a) Toy example of 3 tracklets which could be wrongly linked. The dashed red arrows illustrate wrong assignments that are likely to occur without a coarse-to-fine data association. (b) Coarse-to-fine data association given SAM cardinality. Each sub-graph corresponds to the tracklet association problem over tracklet groups of specific cardinalities, denoted by $\mathbf{C}(sam)$ representing the sum of the elements of the SAM feature. The flow variables obtained by solving these sub-problems are used to defined additional transition costs used in the final optimization.

nected. However, it does not account for the fact that people belonging to groups of different cardinalities (number of people in a group) can still share the same SAM feature. An example is shown in Fig. 7.a, where two tracklets belonging to groups of different cardinalities are wrongly connected (indicated in red) due to similar SAM. However, we want to encourage tracklets from groups of similar sizes to be connected together (black arrows). We account for this by proposing a coarse-to-fine data association method.

We cluster tracklets co-occurring at the same time, into different groups based on the social separation. The cardinality of a tracklet denoted by $\mathbf{C}(x_i)$ is the number of people belonging to the group corresponding to the tracklet x_i. We can imagine that if the clustering is perfect and people moved in the same configuration across the entire camera network, it would suffice to link the tracklet groups instead of the tracklets. This would also solve the problem of tracklets being linked across groups of different cardinalities. However, in practical setting, the grouping is not perfect and people break away from groups. Hence, we link the groups of same cardinality and use the links obtained from this group tracking to define additional transition costs. The complete method is explained in the supplementary document. The method is briefly visualized in Fig. 7.b.

7. Experiments

7.1. Large-scale evaluation

The data collection campaign helps us conduct various experiments in real life setting with a large and dynamic crowd. In this section, we present a set of experiments to address the forecasting problem given the introduced dataset. We select a subset of cameras in our network and measure the performance of our algorithm to forecast mobility, with only these cameras.

Measurement. Previous works have studied the impact of a given detection and tracking algorithm with respect to detailed statistics such as recall/precision rate, MOTA, MOTP and so on [5]. In this work, we are interested in a crowded setting where only part of the scene is covered by a camera network. Hence, we are more interested in evaluating the correct estimation of the origin and destination of a person entering the camera network. We have limited the monitoring to 14 origins and destinations leading to 196 possible OD-path for a trajectory. We have clustered the cameras into two groups: cameras belonging to OD locations (*i.e.* capturing the beginning or ending of long-term tracks), and cameras in-between these locations. We compute the OD error rate as the percentage of wrong predictions out of the total number of people covered by the camera network.

Ground truth. Since Big Data is collected, it is not realistic to label the millions of trajectories. Therefore, we install a dense network of cameras to reduce the blind spots as much as possible and link tracklets that are only a few centimeters away from each other. The trajectories computed from this dense network is used as a baseline. While the trajectories (and OD) computed from the dense network is not the perfect ground truth, in practice they are less easy and less expensive to obtain than manually annotating trajectories at our scale. The goal of our forecasting algorithm is to reach the performance from the dense network of cameras while using a sparse network.

7.2. OD forecasting

Figure 8 presents the resulting OD error rates for 7 sparse networks of cameras. The evaluation is carried out at several levels of network sparsity, from 0% to 75% of in-between cameras. For instance, networks N4 and N5 use only half of the cameras available in the corridor (see figure 2). The cameras are selected to heuristically minimize the average distance between them at any given sparsity. At a given sparsity, we also evaluate on different camera configurations such as N4 and N5 for 50% sparsity. In average, tracklets from network N1 to N3 are several dozen of meters away from each other, and tracklets from networks N4 to N7 are dozen of meters away from each other. To validate our algorithm, we evaluate the performance of greedy optimization methods against the proposed global one. We measure the impact of using SAM as an additional feature, as well as the impact of modeling the OD prior with coarse to fine tracking.

As expected, the global optimization methods always outperform the greedy methods with and without OD prior. The performance improvement is more than doubled, in

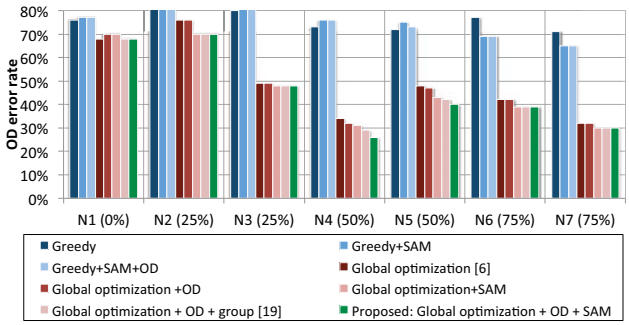


Figure 8: Performance of OD forecasting with different number of in-between cameras. The percentage of in-between cameras are shown in brackets. Seven network configurations are evaluated (referred to as N1 to N7).

the global optimization method. The SAM feature and use of OD-re-weighted cost (use of OD-prior) are both seen to have a positive impact while using global optimization. This justifies our decision to model heuristically model the effect of OD-prior during optimization.

We also compare with the algorithms from [6] and [19]. Our final full model, *i.e*. "Global optimization + OD + SAM", outperforms these methods when observations are limited to the corridor. Note that the camera placement has an impact on the forecasting. Although the same number of cameras are used by networks N2 and N3, or N4 and N5, the forecasting accuracy differs for these networks. If an in-between camera is strategically placed to capture frequent route choices, it reduces the uncertainty in the linking strategy. This leads to different performance for networks with same number of cameras as shown in Fig. 8

We evaluate the extreme setup when there are no in-between cameras (label as N1), *i.e*. we only have cameras at entrance and exit zone (OD cameras). In such setup, tracklets are up to 100 m away from each others. Figure 8 presents the resulting drop in performance. The gap between greedy and global optimization is much smaller. In addition, the SAM feature and OD prior do not have an significant impact on such extreme case. These results motivate our future work to handle such extreme case.

Figure 9 illustrates some qualitative results demonstrating the power of SAM. We also plot the OD prior, forecasted OD with a sparse network of cameras with half the number of cameras as the dense network (ground truth).

Impact of SAM We illustrate the tracklet linking achieved by our full method and compare it with a global optimization method which does not use SAM in Fig. 9. As expected, we see that in the absence of SAM, tracklets travelling in similar group configurations are not connected together, leading to erroneous results. On the other hand, SAM helps disambiguate between tracklet choices which are similar to each other, except for the group configuration.

Impact of OD prior In Fig. 9, we present the final OD-matrices estimated by our full model, and compare it with the OD-prior and the ground truth OD (from dense camera network). Clearly, the prior only provides weak cues about the true OD, but helps by down-weighting paths which are highly unfavorable like blocked corridors. The OD-matrix forecasted by our method is close to the ground truth OD matrix obtained from a dense camera network.

8. Conclusions

We have addressed the problem of forecasting pedestrian destinations with a limited number of cameras in real-world crowded train stations. We have quantitatively shown that social affinities exist and help solve the forecasting problem. The proposed SAM descriptor empowers global optimization of the tracklet association problem with dependent motion behavior. The deployed network of cameras enables a large-scale analysis of real-world crowd motion. Several hundred thousands trajectories are collected per day leading to 42 million trajectories to date. In addition to improving the estimation of the OD matrices, future work can use the data to fine-tune pedestrian simulators, or learn ideal camera placements with a limited number of cameras.

Acknowledgement

We thank P. Vandergheynst, M. Bierlaire, J. Paratte, and D. Chanel for providing useful codes and helpful comments. Alexandre Alahi is funded by the Swiss National Science Foundation under the fellowship no: PBELP2-141078.

References

[1] A. Alahi, L. Jacques, Y. Boursier, and P. Vandergheynst. Sparsity driven people localization with a heterogeneous network of cameras. *Journal of Mathematical Imaging and Vision*, 2011. 1, 2

[2] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *ECCV*. 2008. 2

[3] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR*, 2008. 2

[4] G. Antonini, M. Bierlaire, and M. Weber. Discrete choice models of pedestrian walking behavior. *Transportation Research Part B*, 2006. 2

[5] F. Bashir and F. Porikli. Performance evaluation of object detection and tracking systems. *In PETS*, 2006. 6

[6] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple Object Tracking using K-Shortest Paths Optimization. *TPAMI*, 2011. 1, 7

[7] C. De Vleeschouwer and D. Delannay. Basket ball dataset from the european project apidis, 2009. 2

[8] G. Denina, B. Bhanu, H. T. Nguyen, C. Ding, A. Kamal, C. Ravishankar, A. Roy-Chowdhury, A. Ivers, and B. Varda. Videoweb dataset for multi-camera activities and non-verbal communication. In *Distributed Video Sensor Networks*, pages 335–347. Springer, 2011. 2

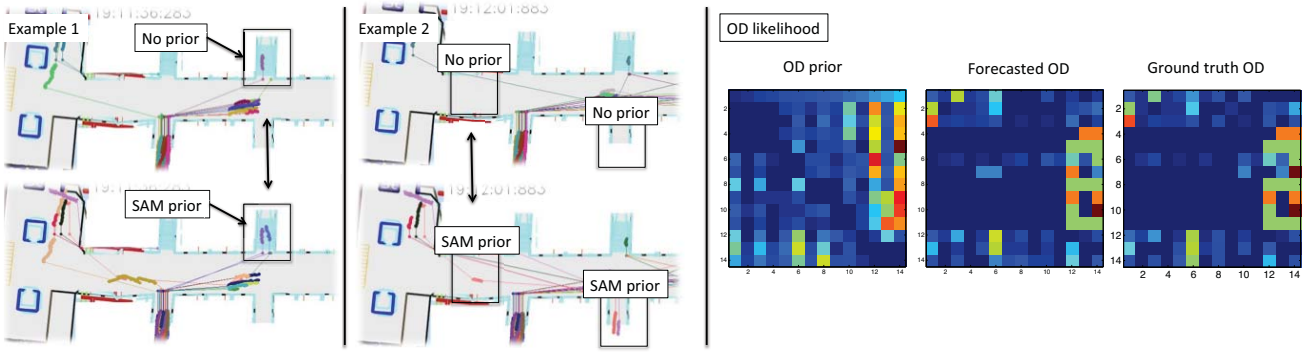


Figure 9: Qualitative results on the linked tracklets within the sparse network 1 where 50% of the in-between cameras within the corridor are not used. Tracklets selected by the method are only shown. The lines illustrate the linked tracklets. On the right side, we illustrate the OD prior as a heatmap, as well as the forecast and ground truth. We can see that although the prior is different, the final result is still similar to the ground truth.

[9] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*. IEEE, 2008. 2

[10] J. Ferryman and A. Ellis. Pets2010: Dataset and challenge. In *Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2010. 2

[11] M. Golbabaee, A. Alahi, and P. Vandergheynst. Scoop: A real-time sparsity driven people localization algorithm. *Journal of Mathematical Imaging and Vision*, 48(1):160–175, 2014. 2

[12] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 1995. 1, 2, 3

[13] S. P. Hoogendoorn and P. H. Bovy. Pedestrian route-choice and activity scheduling theory and models. *Transportation Research Part B: Methodological*, 38(2):169–190, 2004. 1

[14] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. Tracking across multiple cameras with disjoint views. In *Proc. IEEE International Conference on Computer Vision*, page 952, Washington, DC, USA, 2003. IEEE Computer Society. 2

[15] S. M. Khan and M. Shah. Tracking multiple occluding people by localizing on multiple scene planes. *IEEE Transactionson Pattern Analysis and Machine Intelligence*, 31(3):505–519, 2009. 1

[16] K. Kitani, B. Ziebart, J. Bagnell, and M. Hebert. Activity forecasting. 1

[17] C. Kuo, C. Huang, and R. Nevatia. Inter-camera association of multi-target tracks by on-line learned appearance affinity models. *ECCV*, 2010. 2

[18] T. Lan, L. Sigal, and G. Mori. Social roles in hierarchical models for human activity recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[19] L. Leal-Taixe, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *ICCV Workshops*, 2011. 2, 7

[20] K. Lindveld. *Dynamic OD matrix estimation: a behavioural approach*. 2003. 1

[21] M. Luber, J. Stork, G. Tipaldi, and K. Arras. People tracking with human motion predictions from social forces. In *ICRA*, pages 464–469, 2010. 2

[22] 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. 1, 3

[23] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009. 2

[24] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *ECCV*. 2010. 2

[25] A. A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu. Multi-object tracking through simultaneous long occlusions and split-merge conditions. In *CVPR*, 2006. 2

[26] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 2011. 2, 4, 5

[27] Z. Qin and C. R. Shelton. Improving multi-target tracking via social grouping. In *CVPR*. IEEE, 2012. 2

[28] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert. Density-aware person detection and tracking in crowds. In *ICCV*, 2011. 2

[29] Z. Shuai, S. Yoon, S. Oh, and M.-H. Yang. Traffic modeling and prediction using sensor networks: Who will go where and when? *ACM Transactions on Sensor Networks (TOSN)*, 2012. 2

[30] B. Song, T. Jeng, E. Staudt, and A. Roy-Chowdhury. A stochastic graph evolution framework for robust multi-target tracking. *ECCV*, 2010. 2, 5

[31] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, pages 1018–1021, 2010. 1

[32] B. Yang, C. Huang, and R. Nevatia. Learning affinities and dependencies for multi-target tracking using a crf model. In *CVPR*, 2011. 2

[33] Q. Yu, G. Medioni, and I. Cohen. Multiple target tracking using spatio-temporal markov chain monte carlo data association. In *CVPR*, pages 1–8, 2007. 2, 5

[34] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*, 2008. 2, 4, 5