Privacy Preserving Multi-target Tracking

Anton Milan[1] Stefan Roth[2] Konrad Schindler[3] Mineichi Kudo[4]

[1]School of Computer Science, University of Adelaide, Australia
[2]Department of Computer Science, TU Darmstadt, Germany
[3]Photogrammetry & Remote Sensing, ETH Zürich, Switzerland
[4]Division of Computer Science, Hokkaido University, Sapporo, Japan

Abstract. Automated people tracking is important for a wide range of applications. However, typical surveillance cameras are controversial in their use, mainly due to the harsh intrusion of the tracked individuals' privacy. In this paper, we explore a privacy-preserving alternative for multi-target tracking. A network of infrared sensors attached to the ceiling acts as a low-resolution, monochromatic camera in an indoor environment. Using only this low-level information about the presence of a target, we are able to reconstruct entire trajectories of several people. Inspired by the recent success of offline approaches to multi-target tracking, we apply an energy minimization technique to the novel setting of infrared motion sensors. To cope with the very weak data term from the infrared sensor network we track in a continuous state space with soft, implicit data association. Our experimental evaluation on both synthetic and real-world data shows that our principled method clearly outperforms previous techniques.

1 Introduction

Tracking multiple people in indoor environments has many important applications, including customer behavior analysis in retail, crowd flow estimation for building design and planning of evacuation routes, or assistance in daily living for elderly people. While standard surveillance cameras can be employed to address this task, they also have several disadvantages. First, depending on the exact setup, a camera network may be too costly to install and to maintain. Second, standard RGB cameras are highly sensitive to lighting changes and do not work in dark environments. Finally, and most importantly, surveillance cameras are often seen as an intrusion into a person's privacy because they enable a clear identification of the observed person and provide rich visual information about the appearance, pose and exact action of each subject [4, 5].

In this paper we present an affordable, privacy-preserving alternative to address multi-target tracking. To that end, we employ a network of infrared motion sensors that are attached to the ceiling in an indoor scenario. Each sensor is activated whenever a person passes underneath it, yielding a set of sparse measurements that is then used to infer the exact location of each target. To reconstruct the individual trajectories we rely on recent advances in multiple object tracking, *e.g.*, [1–3]. Although the sensory system mounted overhead does not suffer from occlusion, a number of other challenges must be addressed. First, the number of sensors is considerably lower than the number of pixels in a video, leading to a very sparse signal that provides a rather crude approximation of true target locations. Second, a binary sensor response is the only available source of evidence about the presence of a target. Therefore, high-level cues such as a person's appearance or a continuous-valued likelihood of an object detector cannot be exploited. Third, each sensor can be simultaneously activated by several targets, while one single target can activate multiple neighboring sensors when passing between them. Hence we also have to allow many-to-one and one-to-many assignments. Note that the majority of multi-target tracking approaches cannot be directly applied to the present setting because of their common implicit assumptions that each measurement may originate from at most one target and that each target can cause at most one measurement. The strategy we present here is able to handle both cases. Our main contribution is twofold:

- We introduce a novel infrared tracking dataset including the measurements and manually annotated ground truth. The dataset consists of three synthetic and three real world sequences, covering various levels of difficulty.
- We demonstrate how a recent multiple target tracking approach developed for regular cameras [2] can be adopted to address the challenges in this novel setting.

In contrast to previous work in the realm of infrared-sensor tracking [6, 7], we present a simple and more robust tracking method and evaluate its performance using standard tracking metrics. To the best of our knowledge, this is the first time that a global tracking approach is applied to infrared sensor responses. Experimental results show the superiority of our method on several real-world sequences. We make all our data as well as the source code publicly available.[1]

2 Related Work

Research on the automated tracking of multiple targets originated several decades ago in the realm of aerial and naval navigation with radar and sonar sensors. Some of the most notable early works include the multiple hypothesis tracker (MHT) [8] and the joint probabilistic data association (JPDA) [9], which are only rarely used nowadays, as their computation times scale exponentially with the number of targets; these methods hence quickly reach their limits in crowded environments. Such strategies usually apply filtering techniques, for example the Kalman filter [10], in order to estimate the true target locations from noisy observations. More recent approaches follow an offline strategy, where a batch of frames is analyzed at once [1–3, 11–13]. The main motivation behind this is that potential errors may be corrected once more observation steps are available, making these methods more robust against localization noise, false measurements, and target drift.

[1] `http://research.milanton.net/irtracking`

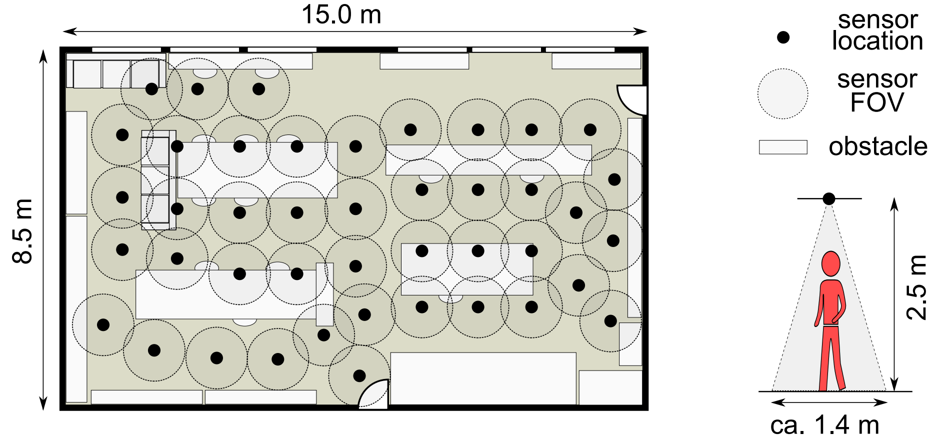

Fig. 1. Overview of our infrared ceiling sensor network.

While RGB cameras or radar/sonar equipment have been the typical sensor choice in the past, tracking results in the literature based on infrared sensors are rather limited. The scheme proposed by Luo *et al.* [7] utilizes a Kalman filter to estimate the location of a single target. However, a rather complex hardware array, where each node consists of five individual sensors equipped with specialized Fresnel lenses, is employed in their setup. Unfortunately, only synthetically generated simulation results are presented. The tracking algorithm described by Hosokawa *et al.* [6] relies on a more complex target localization scheme and includes several ad-hoc procedures to resolve ambiguities. Tao *et al.* [14] also follow a similar setup, but concentrate on activity recognition, in particular fall detection, rather than on tracking individuals in a multi-person scenario.

In contrast, we present an affordable and flexible framework with minimal calibration effort that allows us to robustly keep track of several individuals in an indoor scenario while preserving the individuals' privacy. Quantitative and qualitative evaluation on both synthetic and real-world data demonstrates encouraging results.

3 Infrared Ceiling Sensor Network

We build on the sensor network originally proposed by Hosokawa *et al.* [6]. The entire network consists of 43 nodes attached to the ceiling in a large room of approximately 15.0 × 8.5 meters (*cf*. Fig. 2). Each node is a *pyroelectric infrared sensor*, often simply referred to as an *infrared motion sensor*. The sensor is activated whenever it detects an abrupt temperature change within its range. We exploit this behavior to detect a person moving underneath it. To obtain measurements that are more precisely localized, the detection cone is narrowed to about a 70cm radius on the ground. The sensors are distributed across the entire

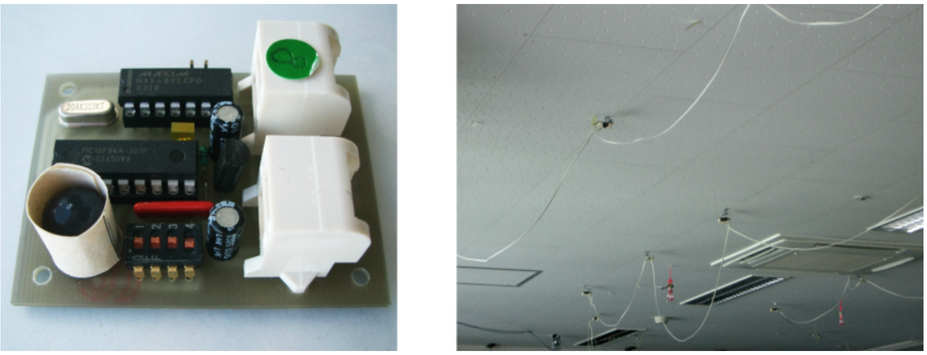

Fig. 2. A sensor node *(left)* and the entire setup *(right)*.

room such that they cover most of the area of interest and their fields-of-view do not overlap substantially (*cf*. Fig. 1). The nodes do not have to be perfectly aligned or arranged in a specific way during installation. An approximate location of each node is sufficient for an accurate calibration, which makes the deployment of such a system rather easy. Note that there is no other information available, such that disambiguating the identity based on visual features is infeasible. The sampling rate of the sensors can be adjusted for a specific application and is set to 2 Hz in our setting. The cost of each sensor is as low as few US Dollars.

4 Multi-target Tracking

Most modern multi-target tracking approaches follow the so-called tracking-by-detection strategy [2, 3, 12, 15, 16], which we also adopt here. In this two-stage strategy a set of measurements is first obtained for each frame independently, forming the target hypotheses. These observations, which are prone to noise and potentially contain false measurements, then serve as input to a tracking algorithm. Moreover, our method also belongs to the class of off-line (non-recursive) state estimation techniques, where, instead of processing one frame at a time as done, *e.g.*, in particle filtering [17], a larger time interval is analyzed in one step. This significantly improves robustness and does not pose a serious limitation, since a slight delay of only a few seconds (in our case 20 frames) is acceptable in practice for the potential applications of our system.

We formulate tracking as minimization of a continuous energy function, which we argue is particularly appropriate for the novel setting that we address here. In particular, we follow the recent work of Milan *et al.* [2] and demonstrate how it can be adapted to this rather different kind of imagery. The state vector $\mathbf{X}$ consists of all (X, Y) coordinates of all targets on the ground plane. We will denote a location of person i in frame t with $\mathbf{X}_i^t$, and the location of the sensor node g with $\mathbf{S}_g$. The set of active nodes at time t is denoted $G(t)$. Finally, N is the total number of targets, and s_i and e_i mark the temporal start and end points of each target, respectively. The energy

$$E = E_{\text{det}} + aE_{\text{dyn}} + bE_{\text{exc}} + cE_{\text{per}} + dE_{\text{reg}}, \tag{1}$$

consisting of a data term, three physically-based (soft) constraints, and a regularizer is then minimized in order to find a locally optimal solution. The approach offers two important advantages. First, trajectories are reconstructed in continuous space such that the low spatial resolution of the sensor network is mitigated. Second, data association is only solved implicitly and not restricted to one-to-one correspondence between observations and target locations. In other words, it is possible that the same measurement may in fact originate from two separate targets, and that one single target can activate two sensors simultaneously. Both situations frequently occur in the observed data and are correctly captured by our model. The individual components are defined as follows:

Observation. The observation term E_{det} keeps the resulting trajectories close to the obtained measurements. To reflect the localization uncertainty of the infrared sensors, we use an inverse Cauchy-like function

$$E_{\text{det}}(\mathbf{X}) = \sum_{i=1}^{N} \sum_{t=s_i}^{e_i} \left[\lambda - \sum_{g \in G(t)} \frac{s^2}{\|\mathbf{X}_i^t - \mathbf{S}_g\|^2 + s^2} \right], \tag{2}$$

where s controls the size of the lobe. Given that the sensors' field-of-view covers a circular area of approximately 1.4 meters in diameter, we employ this value in all our experiments. A uniform penalty λ is applied to all targets to prevent false trajectories without measurements nearby.

Dynamics. The data acquired by infrared sensors is rather limited and exceedingly noisy. A dynamic model is therefore important to bridge missing observations and to restrict data association to plausible solutions. Here, we rely on a constant velocity assumption, and penalize acceleration using

$$E_{\text{dyn}}(\mathbf{X}) = \sum_{i=1}^{N} \sum_{t=s_i+1}^{e_i-1} \|\mathbf{X}_i^{t+1} - 2\mathbf{X}_i^t + \mathbf{X}_i^{t-1}\|^2. \tag{3}$$

Exclusion. Accurately modeling target exclusion is important for several reasons. On one hand, it is desirable to obtain a physically plausible solution without inter-target collisions. On the other hand, we must take into account that a sensor response may be caused by more than one target and that one single target can activate more than one sensor. A continuous exclusion term

$$E_{\text{exc}}(\mathbf{X}) = \sum_{i \neq j} \sum_{t=\max\{s_i,s_j\}}^{\min\{e_i,e_j\}} \frac{1}{\|\mathbf{X}_i^t - \mathbf{X}_j^t\|^2} \tag{4}$$

that directly penalizes situations when two targets come too close to one another serves this purpose. It pushes two trajectories away from one another just enough to avoid a collision, but not too far, since a single measurement should be allowed to explain two targets.

Persistence. Assuming that targets cannot appear or disappear in the middle of the tracking area, the term

$$E_{\text{per}}(\mathbf{X}) = \sum_{\substack{i=1,\dots,N \\ t\in\{s_i,e_i\}}} \frac{1}{1+\exp\big(-q\cdot b(\mathbf{X}_i^t)+1\big)} \tag{5}$$

enforces persistent trajectories and reduces the number of fragmentations. The parameter $q = 1/35$ cm controls the entrance margin and $b(\cdot)$ computes the distance of a trajectory to the border of the tracking area.

Regularization. Finally, we add a regularizer to keep the number of total targets low and enforce longer trajectories:

$$E_{\text{reg}}(\mathbf{X}) = N + \mu \sum_{i=1}^{N} |F(i)|^{-1} \, , \tag{6}$$

where $F(i) := e_i - s_i + 1$ is the total life span of the i^{th} trajectory.

4.1 Optimization

Each energy component is differentiable in closed form, making the entire formulation well suited for gradient-based minimization. We choose to apply a standard conjugate gradient descent to minimize Eq. (1) locally. However, given the highly non-convex nature of the energy, a purely gradient-based optimization would be very susceptible to initialization. Therefore, we add a set of jump moves, as in [2]. These non-local jumps in the energy landscape change trajectory lengths and potentially the number of targets, thus allowing a more flexible probing of the solution space to escape weak minima. Upon convergence of the gradient descent, one of six jump moves described below is executed in a greedy fashion; then the gradient descent restarts.

Growing and shrinking. Each trajectory can be extended by linear extrapolation for an arbitrary number of time steps both forward and backward in time. Similarly, a track is shortened by discarding a fragment of a certain length from either end. Growing is useful for finding new targets, while shrinking weeds out false positives that may have been introduced by noise or during intermediate optimization steps.

Merging and splitting. Two existing trajectories are merged into one if the merge lowers the energy. Note that the individual energy components, in particular the dynamics and the exclusion terms, assert that this step will not cause physically implausible situations with intersecting trajectories or unlikely motion patterns. A single track may also be split into two at a specific point in

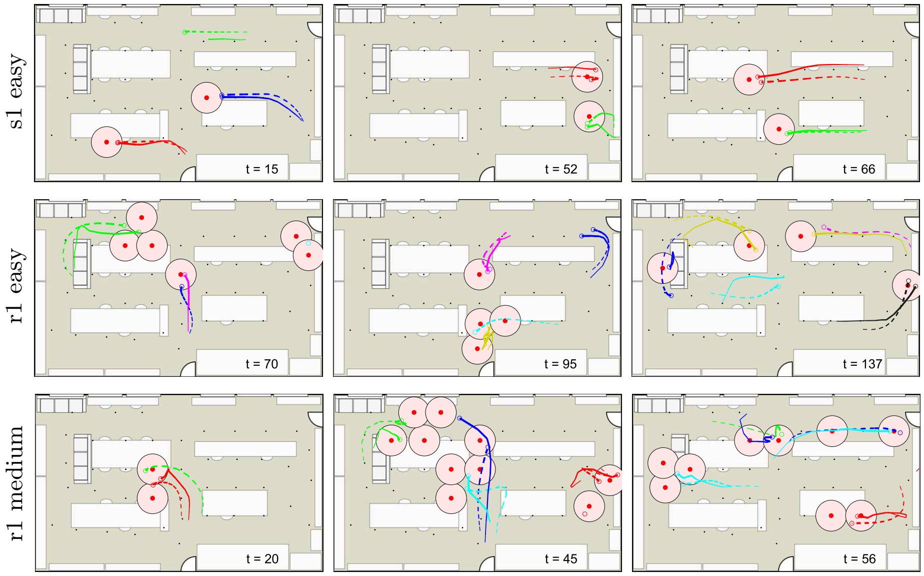

Fig. 3. Qualitative results on synthetic (top row) and real data (middle and bottom rows). 10 frames of both the recovered trajectories (solid) and the ground truth (dashed lines) are shown at three example time steps for each sequence. Note that despite the extreme amount of noise present in the observations (large circles), our method is able to successfully recover most of the targets' trajectories.

time. Both these moves provide a method to bridge over regions with missing sensor responses and to reduce fragmentation of tracks and identity swaps.

Adding and removing. These two moves operate on entire trajectories. Removing a false positive target from the current solution may decrease the overall energy because it results in a more plausible explanation of the data. On the other hand, it is important to allow for inserting new tracks around active sensor locations that do not have a target nearby. This is done conservatively by adding a short tracklet of only three frames. Note that it can grow and merge with other existing trajectories at a later optimization step.

5 Experiments

We present an experimental evaluation of our method on both synthetic and real-world data. Quantitatively assessing the performance of multi-target tracking is an inherently challenging task [18]. Here, we follow the most common strategy and present the evaluation using a set of standard metrics. Next to recall and precision, we compute the *CLEAR MOT* [19] metrics consisting of *MOTA* (Multiple-Object Tracking Accuracy) and *MOTP* (Multiple-Object Tracking Precision). The MOTA includes all possible error types – spurious trajectories or false positives *(FP)*, missed targets or false negatives *(FN)*, and mismatches or identity switches *(ID)* – and is normalized such that 100% corresponds to no errors. The MOTP directly measures the performance of location estimation by computing the average distance between the true target and the inferred location, again normalized to 100%. We use a 1.5m hit/miss threshold on the ground plane. The weights a through d for the individual energy terms in Eq. (1) were determined empirically and kept fixed for all experiments at $\{.0006, .8, .08, .02\}$. The additional parameters λ and μ were set to .004 and 1, respectively.

5.1 Datasets

Synthetic data. Acquiring large amounts of accurate ground truth for multi-target tracking is tiresome and costly. Therefore, we first test our presented method on a synthetic dataset. The data is created by simulating plausible trajectories and the generated sensor responses. Trajectories are spline interpolations between sparse key points corresponding to typical motion patterns. An average target speed of 1m/s with Gaussian noise is assumed for our purpose. A sensor is set to 'active' if at least one trajectory passes within a distance close than its range of operation, which amounts to 70cm. Three sequences (*s1 easy/medium/hard*) with two, four and six targets, respectively, present a reasonable variability in person count and density.

Real-world data. While simulated data may be comparatively easy to acquire, it typically does not fully capture the complexity of real observations. Thus, it is essential to also test a system on real sensor data. To that end, we recorded three sequences of approximately two minutes each. Six persons were moving freely around the entire walkable space inside the lab. The ground truth was annotated manually, relying on videos from two cameras that served as reference (*cf*. Fig. 4). Note that a precise localization of each person is ambiguous – and sometimes even impossible – with the available setup, due to low resolution and/or occlusions. Nonetheless, this new dataset is a sensible basis to quantitatively evaluate the tracking performance. The entire dataset consists of 974 frames, which amounts to over eight minutes of data. The average person count across all sequences is 3.2.

Qualitative results of our proposed method are shown in Figure 3. Each row depicts three frames from one particular sequence, while in each frame the currently active sensors are indicated with large circles, and the estimated and true trajectories are plotted with solid, respectively dashed lines for the past ten time steps. Note the extremely noisy observations, including spurious activations and missing signal.

Our method is able to correctly recover the number of targets and their trajectories in cases where the targets remain well separated (see top row). Note that a precise localization is not always possible due to the scarcity of the available data, as can be seen, *e.g.*, for the red target. In realistic environments with more targets and sensor noise, there is a certain drop in performance, as expected. The

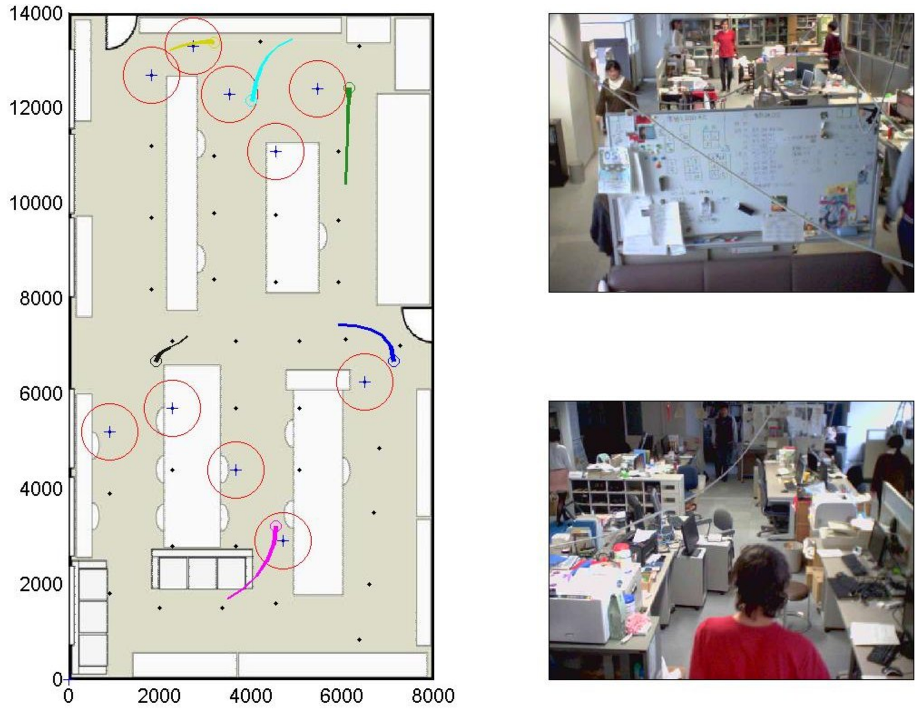

Fig. 4. A screenshot of the data acquisition setup. The video captured by two cameras on the right serves as reference for annotating each individual's location on the ground plane. The large red circles indicate active sensors in the current frame.

main cause of failure are from ambiguous measurements, which typically occur when people walk close to each other for a prolonged period of time (*cf*. bottom left frame). In such cases, the sensors are unable to provide enough information to robustly resolve the ambiguity.

5.2 Quantitative Evaluation

Table 1 shows quantitative results of our proposed method on all six sequences (three synthetic (s*) and three real (r*) ones). Note that we achieve near perfect precision and only few identity switches with simulated data, i.e. in the absence of sensor noise. The performance decreases for real data, but still stays above 50% MOTA on average. Table 2 lists average results of our method compared to those of three other strategies. One is a linear location estimation technique specifically designed for inrared motion sensors, where the number of targets is determined by connected components of sensor responses [14].[2] While it can recover most targets without producing too many false positives, the number of

[2] Implementation provided by the authors.

Table 1. Quantitative results on synthetic *(top)* and real *(bottom)* data.

Sequence	Recall	Precision	MOTA	MOTP	FP	FN	ID
s1 easy	93.3 %	98.9 %	88.2 %	76.6%	2	13	8
s1 medium	82.3 %	97.1 %	76.2 %	73.1%	9	64	13
s1 hard	71.0 %	94.3 %	63.5 %	71.0%	25	170	19
mean (s*)	**82.2 %**	**96.8 %**	**76.0 %**	**73.6%**	**12**	**82**	**13**
r1 easy	84.2 %	81.0 %	57.5 %	53.8%	143	114	50
r1 medium	74.5 %	83.2 %	52.9 %	56.9%	88	149	38
r1 hard	74.4 %	85.3 %	55.4 %	53.1%	86	172	42
mean (r*)	**77.7 %**	**83.2 %**	**55.3 %**	**54.6%**	**106**	**145**	**43**

Table 2. Comparison to other methods averaged over synthetic *(top)* and real *(bottom)* sequences. The best average performance for each measure is highlighted in bold face.

	Method	Recall	Precision	MOTA	MOTP	FP	FN	ID
synth.	Linear [14]	81.0 %	**99.8 %**	66.6 %	64.6 %	**1**	**81**	58
	DP [20]	78.5 %	92.0 %	55.9 %	65.3 %	27	81	57
	KSP [12]	78.9 %	97.5 %	75.5 %	67.5 %	6	83	**6**
	Ours	**82.2 %**	96.8 %	**76.0 %**	**73.6 %**	12	82	13
real	Linear [14]	79.3 %	71.5 %	9.3 %	50.1 %	212	137	252
	DP [20]	71.6 %	62.7 %	9.6 %	47.3 %	281	188	128
	KSP [12]	**89.4 %**	63.7 %	31.1 %	48.3 %	337	**70**	48
	Ours	77.7 %	**83.2 %**	**55.3 %**	**54.6 %**	**106**	145	**43**

ID switches is quite high. The second baseline is the globally optimal approach by Pirsiavash *et al.* [20] based on dynamic programming (DP). It is able to better keep correct identities over time, but struggles to handle the extremely noisy measurements leading to a high number of false alarms and missed targets. Finally, the third method is the k-shortest paths (KSP) approach [12], where targets are tracked on a discrete grid. Its power to robustly keep target identities over long time periods is unfolded in the present setting with overhead sensors and in absence of occlusion, particularly with noiseless synthetic data. However, the coarse discretization of the grid is not able to handle real-world noisy measurements, leading to many false tracks. The continuous state representation in combination with the soft assigment strategy of our method clearly outperforms previous techniques with respect to the most relevant tracking metrics (MOTA, MOTP).

Finally, Table 3 shows the mean average people count error of the four methods, again split into two groups of synthetic and real data. Although all four methods show similar performance in the absence of noise, our proposed ap-

Table 3. Person count estimation. Per-frame mean absolute error (MAE) and standard deviation is shown for synthetic and real data.

Method	Linear [14]	DP [20]	KSP [12]	**Ours**
MAE (synth.)	0.57±0.78	0.62±0.75	0.57±0.75	**0.54±0.81**
MAE (real)	1.00±0.95	1.25±1.16	1.52±1.06	**0.76±0.76**

proach provides the most accurate estimate on the number of people present in the scenes with realistic data.

6 Discussion and Conclusion

We have presented a method for tracking multiple people while fully preserving their privacy. A ceiling infrared sensor network serves as a low resolution camera providing only a sparse binary signal about the presence of moving targets. Building on recent advances in multiple target tracking, we follow an energy minimization strategy to localize walking people and reconstruct their trajectories in spite of the impoverished observation data. Our framework outperforms other methods measured with respect to widely used tracking metrics, both on synthetic and real-world data.

The low spatial resolution and high noise level of the signal provided by the sensors clearly limits the ability to resolve all identity ambiguities. A precise localization as well as targets that remain still for longer time periods causing a sensor to become inactive still remain challenges that should be addressed in future work.

References

1. Jiang, H., Fels, S., Little, J.J.: A linear programming approach for multiple object tracking. (In: CVPR 2007)
2. Milan, A., Roth, S., Schindler, K.: Continuous energy minimization for multitarget tracking. IEEE T. Pattern Anal. Mach. Intell. **36** (2014) 58–72
3. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. (In: CVPR 2008)
4. Babaguchi, N., Koshimizu, T., Umata, I., Toriyama, T.: Psychological study for designing privacy protected video surveillance system: PriSurv. In Senior, A., ed.: Protecting Privacy in Video Surveillance. Springer London (2009) 147–164
5. Norris, C., Armstrong, G.: CCTV and the social structuring of surveillance. In: Surveillance of Public Space. Volume 10 of Crime Prevention Studies. Criminal Justice Press (1999) 157–178
6. Hosokawa, T., Kudo, M., Nonaka, H., Toyama, J.: Soft authentication using an infrared ceiling sensor network. Pattern Anal. Appl. **12** (2009) 237–249
7. Luo, X., Shen, B., Guo, X., Luo, G., Wang, G.: Human tracking using ceiling pyroelectric infrared sensors. In: IEEE International Conference on Control and Automation, 2009. ICCA 2009. (2009) 1716–1721

8. Reid, D.B.: An algorithm for tracking multiple targets. IEEE Transactions on Automatic Control **24** (1979) 843–854
9. Fortmann, T.E., Bar-Shalom, Y., Scheffe, M.: Multi-target tracking using joint probabilistic data association. In: 19th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes. Volume 19. (1980) 807–812
10. Kalman, R.E.: A new approach to linear filtering and prediction problems. Transactions of the ASME–Journal of Basic Engineering **82** (1960) 35–45
11. Andriyenko, A., Schindler, K., Roth, S.: Discrete-continuous optimization for multi-target tracking. (In: CVPR 2012)
12. Berclaz, J., Fleuret, F., Türetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. IEEE T. Pattern Anal. Mach. Intell. **33** (2011) 1806–1819
13. Butt, A.A., Collins, R.T.: Multi-target tracking by lagrangian relaxation to min-cost network flow. (In: CVPR 2013)
14. Tao, S., Kudo, M., Nonaka, H.: Privacy-preserved behavior analysis and fall detection by an infrared ceiling sensor network. Sensors **12** (2012) 16920–16936
15. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. (In: CVPR 2008)
16. Zamir, A.R., Dehghan, A., Shah, M.: GMCP-Tracker: Global multi-object tracking using generalized minimum clique graphs. In: ECCV 2012. (Volume 2.) 343–356
17. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Robust tracking-by-detection using a detector confidence particle filter. (In: ICCV 2009)
18. Milan, A., Schindler, K., Roth, S.: Challenges of ground truth evaluation of multi-target tracking. (In: Proc. of the CVPR 2013 Workshop on Ground Truth - What is a good dataset?)
19. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The CLEAR MOT metrics. Image and Video Processing **2008** (2008) 1–10
20. Pirsiavash, H., Ramanan, D., Fowlkes, C.C.: Globally-optimal greedy algorithms for tracking a variable number of objects. (In: CVPR 2011)