Boston U. Computer Science Tech. Report No. BUCS-TR-2008-008, June 2008.
To appear in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, **June 2008.**

Layered graphical models for tracking partially-occluded objects

Vitaly Ablavsky, Ashwin Thangali, and Stan Sclaroff
Computer Science Department, Boston University, USA
{ablavsky, tvashwin, sclaroff}@cs.bu.edu

Abstract

Partial occlusions are commonplace in a variety of real world computer vision applications: surveillance, intelligent environments, assistive robotics, autonomous navigation, etc. While occlusion handling methods have been proposed, most methods tend to break down when confronted with numerous occluders in a scene. In this paper, a layered image-plane representation for tracking people through substantial occlusions is proposed. An image-plane representation of motion around an object is associated with a pre-computed graphical model, which can be instantiated efficiently during online tracking. A global state and observation space is obtained by linking transitions between layers. A Reversible Jump Markov Chain Monte Carlo approach is used to infer the number of people and track them online. The method outperforms two state-of-the-art methods for tracking over extended occlusions, given videos of a parking lot with numerous vehicles and a laboratory with many desks and workstations.

1. Introduction

In computer vision, reliable tracking through occlusions is a critical problem. Even partial occlusions can confound many trackers, leading to fragmentation or total loss of tracks. While inference methods have been proposed that can stitch track segments together and link across gaps, such methods can be brittle. And while detailed 3D scene models can be used to predict and handle occlusions, such models must be painstakingly defined for each scene.

This paper proposes a framework for tracking multiple people who move around in structured environments with numerous occluders, like desks in an office, or vehicles in parking lots. The image-plane representation of motion around each object is associated with a pre-computed graphical model, which can be instantiated efficiently during online tracking. A global state and observation space are obtained by linking transitions between layers. In our experiments, the method outperforms two state-of-the-art methods [7, 8] in tracking over extended occlusions.

2. Basic Idea

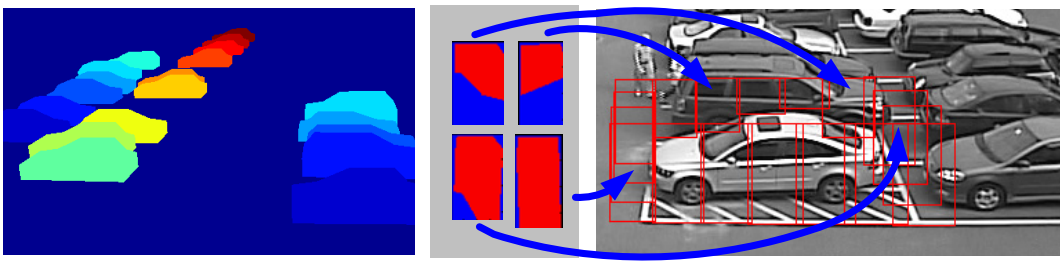
In the proposed approach, motion patterns around a known object, for instance a desk or a vehicle, are first abstracted by considering zones around that object in 3D. An example parking lot application is shown in Fig. 1. Zones are object-centered in the sense that a zone location for "emerging from driver's side of car" is generic in a car-centered coordinate system, but its location in world coordinates depends on the car's position and orientation in the world. For person tracking applications, each activity zone is a person-sized box in 3D.

Transitions between an object's zones are represented via a Markov model: each zone has a vertex in the model, and edges convey the degree to which a zone is "reachable" from another. For instance, movements between adjacent zones around a car are probable, but jumps over a car are not. A single graphical model can be trained for a whole class of objects and then instantiated as needed by copying the graphical model for each object instance. Edges between the objects' graphical models are determined by the proximity of corresponding zones in world coordinates.

Pixels corresponding to a particular zone's projection in the image plane are approximated by a rectangle, hereafter referred to as a *receptive field*. We define a 2.5D graphical model comprising (a) the occluder's mask (b) a set of receptive fields depth-ordered with respect to and clipped by the mask (c) a state graph whose nodes are identified with receptive fields and whose edges encode transition probabilities. Multiple 2.5D models interact when their masks overlap, as shown in Fig. 2 and described in detail in Sec. 4.

Given approximate ground plane calibration, a database of such 2.5D models can be pre-computed. Subsequently, we can instantiate our representation, given image-plane locations in partial order of model instances. As our experiments show, the proposed framework can handle coarse masks (e.g., for modelling cars, only two categories: sedan and van), which are not exactly aligned.

Receptive fields can overlap in the image as shown in Fig. 1(c). The challenge then is to infer the active zone(s) based on motion observed in overlapping receptive fields.



(a) (b) (c)

Figure 1. Illustrative example: parking lot application. (a) 3D vehicle models are placed in the scene given ground plane calibration, and car masks are rendered via computer graphics. Each car mask instantiates a layer of the graphical model, as described in the text. (b) Occlusion of some receptive fields by the depth-ordered vehicle masks (red pixels are not occluded). (c) Receptive fields obtained as projections of person-sized activity zones surrounding a vehicle, where receptive fields take into account occlusion from car masks in front.

Given the potentially huge number of zones (thousands), it is impractical to define a state-space dimension equal to the number of zones. Instead, the dimension can be defined as the number of people in the scene at any given time. This state-space dimension will increase/decrease as people enter/leave the scene. Reversible Jump Markov Chain Monte Carlo provides an efficient framework for approximate inference and tracking in this setting [7].

Our approach enables tracking of people moving among objects, with multiple partial occlusions, without losing track. The occlusion handling represents an improvement over prior work, as outlined in the next section.

3. Related Work

The BraMBLe system [2] utilized a 3D representation: knowledge of the camera geometry was combined with a generalized-cylinder representation of a human to model the foreground accurately. By utilizing the probabilistic exclusion principle the system effectively reasoned about the number and location of multiple occluding people in a hallway scene. Such a design is suitable for short-term and self-occlusions, but not for situations where portions of people are invisible for extended periods of time.

Many methods have been proposed that reason about occlusion in the image plane, operating purely in 2D. Handling of multi-object occlusions in person tracking was studied by Pérez, et al. [5], but experiments were limited to scenarios with short-lived occlusions. An appearance-based tracker that handles short-term occlusions was proposed by Takala and Pietikainen [8], but this method does not compare favorably in our experiments.

In another 2D approach, Smith, et al. [7] propose a Reversible Jump Markov Chain Monte Carlo (RJMCMC) based particle filter to accommodate shrinking and expansion of the state space in tracking varying number of people. A global observation model including both foreground and background distributions enables comparison of likelihoods for different numbers of hypothesized pedestrians to help accurately predict the number of people. A Markov Random Field (MRF) via an interaction potential term keeps clusters of nearby image rectangles from collapsing to the same location. A similar interaction term is also employed by Khan, et al. [4] for tracking ants.

The framework proposed in this paper belongs to a class of 2.5D approaches that employ layered scene representations to handle static occluders. Learning of occlusion masks and their depth ordering has been demonstrated by Renno, et al. [6]. Assuming pedestrians take all possible paths through the scene, the system can learn per-pixel integer depths of static objects. However, the requirement that pedestrians explore a large portion of the scene before static occluders are resolved can be limiting. For instance, in a parking lot surveillance application occluders (cars) arrive and leave frequently, but the occlusion map must be available immediately for pedestrian tracking.

Tracking through occlusions via a set of foreground and background layers was proposed by Zhou and Tao [11]. However, that framework did not model motion of objects *around* occluding layers.

Our sampling-based inference algorithm for tracking is influenced by the work of Smith, et al. [7]. Their formulation is designed for a continuous state space and works well when there are no significant occlusions. In our formulation, people are constrained to move on an object-centered grid allowing occlusion-aware reasoning and structural constraints to be incorporated. These two approaches are hence complementary and we envision a handoff with the former tracking in free space and the latter taking over in regions with known static occluders.

4. Approach

We now elaborate on the basic ideas of the framework given in Sec. 2. For clarity of exposition, the approach is illustrated in a specific application: pedestrian tracking in a parking lot. However, it should be noted that the framework is general and thus can be used in other application settings.

Each *object layer* is a graphical model defined by: an occlusion car mask obtained by rendering a 3D vehicle template, receptive field image regions surrounding the vehicle (their firing will serve as the observations), transition prob-

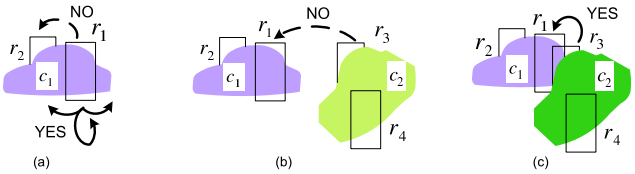


Figure 2. Example interactions between receptive fields. (a) A vehicle centric model allows transition between neighboring zones but not jumping between front and back. (b) Two instantiated activity models that are spatially disjoint do not interact. (c) Spatially overlapping masks induce interaction between corresponding activity models: a person in front of c_1 can transition behind c_2.

abilities between receptive fields, and an observation likelihood for receptive field firing given a pedestrian state. Each object in the scene has its own graphical model layer, and layers are sorted by increasing depth from the camera.

Receptive fields are rectangles that approximate the image projections of person-sized activity zones defined around the vehicle. Fig. 1(c) illustrates the instantiation of receptive field layers for one vehicle in a scene. Receptive fields are cropped by masks of vehicles that come later (closer to the viewer) in the depth-ordered layers; each receptive field hence corresponds only to the visible portion of its activity zone.

Transition probabilities between receptive fields of an object layer are based on the expected motion patterns around an instance of that particular object class. For example, for each vehicle in the parking lot application, a person is more likely to remain stationary near a car door, unlikely to jump over the car, etc.

With all object layers defined, we add interactions between them to obtain a global graphical model for person tracking. Overlap and/or proximity can induce transitions between receptive fields of adjacent vehicles. Fig. 2 illustrates intuitively some of the constraints for receptive field transitions within and between vehicle layers. The resulting global transition matrix shown in Fig. 3 demonstrates the significant sparsity induced by the above constraints.

In the next section, an observation likelihood is formulated taking into account correlations in the receptive field firings due to their overlap in the image plane.

4.1. Observation Likelihood

Consider the observation likelihood $P(Z_t|X_t)$, where $Z_t \in \mathbb{R}^N$ is the observation vector consisting of N receptive field responses and $X_t = \{x_t^1, ..., x_t^k\}$ the set of discrete receptive field indices occupied by k pedestrians at time t. Each pedestrian occupies one receptive field at each frame.

Receptive fields regions in the image overlap; thus, their responses are correlated. A pedestrian hypothesized at a specific activity zone will lead to firing in the corresponding and overlapping receptive fields. To account for this, we define an MRF over hidden variables ν, as depicted in the example of Fig. 4. The ν are considered as random variables influenced by the underlying pedestrian state variable x. The edges (ν_i, ν_j) are only present for receptive fields that overlap, i.e. $r_i \cap r_j \neq \emptyset$. The observations z_i are linked to the corresponding ν_i.

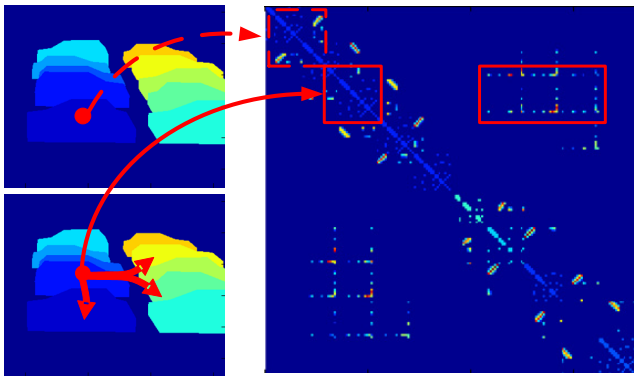


Figure 3. Each instantiated layer model contributes a block to the global transition matrix on the right. For example, the upper-left block (dashed red square) is specified by the activity model of the left-nearest model. The model behind it contributes the second diagonal block. Interaction between activity models is encoded by the off-block-diagonal entries shown in the red rectangle.

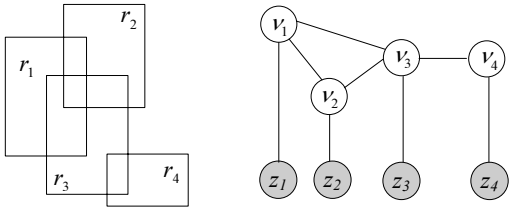


Figure 4. Four receptive fields and their MRF to account for correlations. The ν are considered as random variables influenced by the underlying pedestrian state variable x. The observations z_i are linked to the corresponding ν_i. Edges between ν are only present for receptive fields that overlap.

Given the graph G, we define a global observation model with respect to a set C of maximal cliques c

$$p(z_1, \ldots, z_N, \nu_1, \ldots, \nu_N) \propto \prod_{n=1}^{N} \psi_n(z_n, \nu_n) \prod_{c \in C} \psi_c(\nu_c)$$

where ψ_c is a compatibility function defined to capture correlations. For instance in Fig. 4, $\psi_c(\nu_1, \nu_2, \nu_3)$ assigns low compatibility to $(\nu_1 = 1 - \epsilon, \nu_2 = 1 - \epsilon, \nu_3 = \epsilon)$ for small $\epsilon > 0$. We simplify the model assuming a deterministic relationship between ν and x given by,

$$\nu_n | X = \frac{\text{area}(r_n \cap r_X)}{\text{area}(r_n)}.$$

Where, r_X is the union of receptive field regions corresponding to hypothesized person locations, $r_X = \bigcup_{n \in X} r_n$. The observation likelihood can then

be simplified as,

$$\begin{aligned} p(z_1,\ldots,z_n|X) &= p(z_1,\ldots,z_n|\nu_1,\ldots,\nu_n) \\ &= \prod_{n=1}^{N} p(z_n|\nu_n). \end{aligned}$$

To obtain $p(z_n|\nu_n)$ a distribution linking observed features to overlap values, we employ a linear combination of two learned densities: $p_{\text{in}}(\cdot)$ for observations in a receptive field that fully overlaps a moving pedestrian and $p_{\text{out}}(\cdot)$ for receptive field response not caused by pedestrian motion, i.e., due to image noise:

$$p(z_n|\nu_n) = \nu_n p_{\text{in}}(z_n) + (1-\nu_n)p_{\text{out}}(z_n).$$

In our implementation, we have experimented with two types of observation features.

Motion features. Motion-based features are useful when the distance to objects is large or when occlusions are severe. These features are obtained by averaging motion estimated within each receptive field.

Background subtraction features. When foreground objects occupy a sufficient number of pixels, precision based features can be used [7]. The observation value z_n equals the percentage of the receptive field r_n occupied by the foreground blob.

4.2. Tracking/inference

Given the above observation model, the multi-person state space is given by a Cartesian product of single person states, $X_t = \{x_i\}_{i=1}^{k}$ for $k \geq 0$. Since the state space is large and changes dimensionality when people arrive and depart, exact inference is intractable. Even on our discretized grid, the number of states N^k is enormous.

Reversible-Jump Markov Chain Monte Carlo (RJMCMC) is particularly suited to this problem. The varying state space dimensionality is handled by employing birth and death moves to hypothesize a new track or remove an existing one. Smith, et al. [7] employed this approach with excellent results for tracking pedestrians – but unlike our framework, their system did not handle substantial prolonged occlusions by static objects. We now extend the RJMCMC filter of [7] to discrete state space by reformulating the acceptance ratios for birth, death, and update moves.

For the discrete formulation, the filtering distribution

$$p(X_t|Z_{1:t}) = C^{-1}p(Z_t|X_t) \times \int_{X_{t-1}} p(X_t|X_{t-1})p(X_{t-1}|Z_{1:t-1})dX_{t-1}$$

is approximated by M samples

$$p(X_t|Z_{1:t}) \approx C^{-1}p(Z_t|X_t)p_0(X_t)\sum_{m=1}^{M} p_V(X_t|X_{t-1}^{(n)})$$

$p(Z_t\|X_t)$	observation likelihood at time t
$p_V(X_t\|X_{t-1})$	person dynamics; a random walk with given receptive field transition probabilities
$\{X_t^1,...,X_t^S\}$	chain of MCMC sampled states at time t
X_t^{n*}	a proposed state at iteration n of MCMC chain for time t
X_t^n	accepted state at n^{th} iteration
i^*	id of target chosen at a MCMC iteration to apply one of the moves, $i^* \in \{1,...,k+1\}$
$\alpha_b, \alpha_d, \alpha_u, \alpha_s$	acceptance probabilities for MCMC moves
$p_v(\cdot)$	probability for sampling each of four moves
$\phi(x_i, x_j)$	interaction term to prevent collapse of multiple people states onto a single location
$q_b(\cdot), q_d(\cdot)$	constraints on receptive field locations where pedestrians can enter/exit the scene

Table 1. Symbols for the MCMC sampler formulation.

Samples from Eq. 1 are drawn via MCMC with four move types: *birth*, *death*, *update*, *swap*. *birth* move changes the model order from k to $k+1$, *death* move is its inverse, *update* changes target's position, and *swap* swaps identities for a pair of targets.

In iteration n of the MCMC chain at time t, a state X_t' is chosen at random from the sample set at $t-1$, $X_t' \sim \{X_{t-1}^1,...,X_{t-1}^S\}$. A target $i*$ and move v are randomly chosen and applied to X_t' resulting in a proposed state X_t^{n*}. X_t^{n*} is accepted with probability $\alpha_{(\cdot)}$, $X_t^n = X_t^{n*}$; if rejected, $X_t^n = X_t^{n-1}$.

The *Birth move*'s proposal distribution $q_b(\cdot)$ keeps all current objects fixed and assigns non-zero probability to configurations containing a new target i^*. $\phi(x_i, x_j)$ is an interaction term to prevent states of multiple people from collapsing onto a single location. The acceptance ratio is

$$\begin{aligned} \alpha_b &= \min(1, R_b) \\ R_b &= \frac{p(Z_t|X_t^{n*})}{p(Z_t|X_t^{n-1})} \frac{\prod_{j\in C_{i*}} \phi(X_{i*,t}^{n*}, X_{j,t}^{n*})}{1} \frac{p_v(death)}{p_v(birth)} \frac{q_d(i*)}{q_b(i*)} \end{aligned}$$

The *Death move*'s proposal distribution $q_d(\cdot)$ assigns non-zero probability to configurations in which all objects are fixed and i^* has been removed. The acceptance ratio is

$$\begin{aligned} \alpha_d &= \min(1, R_d) \\ R_d &= \frac{p(Z_t|X_t^{n*})}{p(Z_t|X_t^{n-1})} \frac{1}{\prod_{j\in C_{i*}} \phi(X_{i*,t}^{n-1*}, X_{j,t}^{n-1*})} \\ &\quad \times \frac{p_v(birth)}{p_v(death)} \frac{q_b(i*)}{q_d(i*)} \end{aligned}$$

The *Update move*'s proposal distribution incorporates target dynamics $p_V(\cdot)$ for target i^* while all other targets fixed. The acceptance ratio is

$$\alpha_u = \min\left(1, \frac{p(Z_t|X_t^{n*})}{p(Z_t|X_t^{n-1})} \times \frac{\prod_{j\in C_{i*}} \phi(X_{i*,t}^{n*}, X_{j,t}^{n*})}{\prod_{j\in C_{i*}} \phi(X_{i*,t}^{n-1}, X_{j,t}^{n-1})}\right)$$

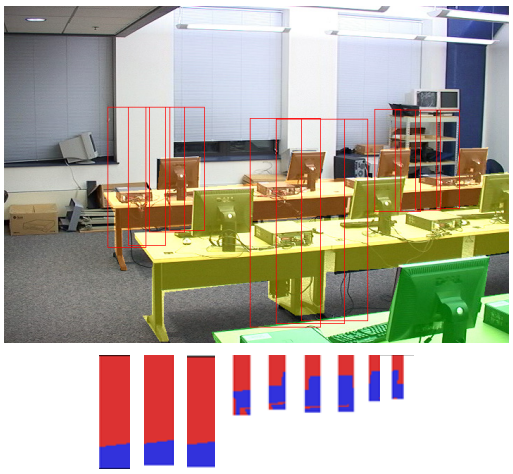


Figure 5. Three rows of tables with computers create occlusion layers. Receptive fields are instantiated for all layers. A few of the receptive fields are displayed (visible parts are shown in red and occluded parts in blue).

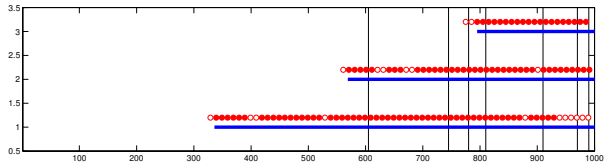


Figure 6. Tracking result time-lines for laboratory sequence whose sample frames are shown in Fig. 7. Vertical axis gives the track-id and horizontal axis the frame numbers. Blue segments are ground-truth start and end for person tracks. The circles in red are tracking results for our approach. Solid circles indicate that the person was hypothesized in the correct layer and an open circles implies an incorrect layer. Vertical lines mark instances of inter-person occlusions. Details are further described in the text.

The *Swap move*'s proposal distribution swaps two targets' state values and histories keeping the rest fixed. The acceptance ratio is

$$\alpha_s = \min\left(1, \frac{p(Z_t|X_t^{n*})}{p(Z_t|X_t^{n-1})}\right)$$

Using the above tracking formulation, it becomes possible to track a time-varying number of people through challenging occlusions. Moreover, this can be done at near video frame rates, as is demonstrated in the next section.

5. Experiments

The method is now demonstrated in tracking multiple people, with occlusions, in two different settings: a computer laboratory with tables and workstations, and an outdoor parking lot in daytime.

5.1. Video of computer laboratory

The first example is the simpler of the two, and is designed to demonstrate the method's ability to assign tracks to appropriate layers. Video sequences were collected in a lab setting, as shown in Fig .5. Three rows of tables with workstations serve as occlusion layers, the aisle on the left is modeled as a transparent layer. The layer masks are instantiated manually. Ground plane calibration is used to instantiate person-sized receptive fields in each layer (total number of receptive fields is 64).

The parameter settings for the implementation of our method were as follows. For the MCMC sampler, probability of birth, death, update, and swap moves were $[0.1, 0.01, 0.9, 0.0]$ (the probability of swaps was set to zero, since there was no model of appearance). The observation likelihood $p_{\text{in}}(z_n)$ was specified using a two-component Gaussian mixture model, and $p_{\text{out}}(z_n)$ was modeled as a single Gaussian. The interaction potential $\phi(\cdot,\cdot)$ was modeled as $\phi(\cdot,\cdot) \propto exp(-\lambda_I\, \mu)$ with μ specifying the ratio of receptive field overlap over their union, $\lambda_I = 100$ for objects in the same layer and $\lambda_I = 0$ for objects in different layers. The observation vector is comprised of precision values for receptive fields, as described in the previous section. Background subtraction is performed using Local Binary Pattern histograms (LBPH) [1, 8]. 40 samples are drawn in the MCMC chain of which 25% are burn-in samples.

Sample frames from the tracking sequence are shown in Fig. 7 and a time-line summary is shown in Fig. 6. As can be seen, the three persons entering the scene are tracked accurately with respect to their image plane locations and their layer assignment. Layers are sometimes times confused when the person is walking in the left aisle since motion here is less constrained than when walking between desks. The tracker can handle many instances of persons occluding each other and occlusions by static objects (desks and workstations) in the scene.

5.2. Videos of parking lot

In the second example, unscripted video sequences were collected during the morning rush hour at an office parking lot. The "MINI-Cooper" sequence (2000 frames) contains two people walking together first in the open area and then amidst parked cars. The "far-field" sequence (2300 frames) contains two drivers getting out of cars while another vehicle passes. A tracker in this setting has to contend with low resolution and prolonged occlusions. Given the repeated structure of static occluders (the cars) this setup is a good match for our approach utilizing pre-computed layers.

Fig. 9(a) shows the performance of Smith's approach [7] for the "MINI-Cooper" sequence. AR0 dynamics (random walk) and the LBPH foreground model are used for fairness of comparison. The tracker is initialized with two pedestrian models in frame 588. The tracker handles brief occlusions of one pedestrian by another (in frame 1129). In frame 1223 only heads are visible as indicated by two green arrows and both tracks are lost. This is to be expected since the appearance and size of the foreground blob

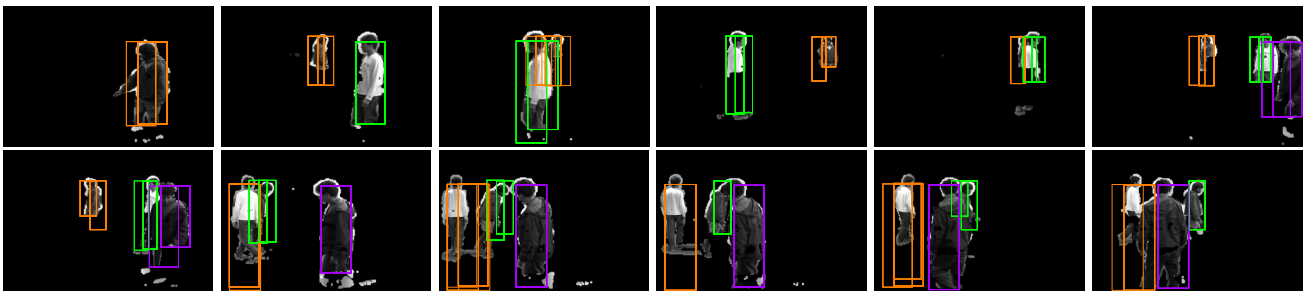


Figure 7. Samples from the multi-person posterior for the indoor lab sequence corresponding to timeline in Fig. 6. The rectangles correspond to receptive fields and are different sizes based on their layer and cropping by occlusion masks in front. The tracker is able to track three persons who enter the scene through multiple instances of occlusion. Layer associations for the most part are correct. Identity swaps do happen since we do not incorporate a person appearance model.

changes drastically. On the second "far-field" sequence, too few pixels are available to learn a stable appearance model leading to unreliable performance.

Fig. 9(b) shows the performance of Takala's tracker [8] on the MINI sequence. In this case, the tracker automatically initializes to track the same two pedestrians initially as a single entity, then splits them correctly into two pedestrians, and correctly detects occlusion of one pedestrian by another. Unfortunately, this tracker also looses track when only heads are visible (indicated by green arrows in the figure). With the "far-field" sequence, Takala's tracker only picks up the track when a person moves close to the camera, the results are not shown due to space limitations.

For our approach, a car layer is pre-computed with its occlusion mask, 18 receptive fields, and associated transition probabilities. Car masks for vehicles in the lot are positioned manually given ground plane calibration. A total of 230 receptive fields are automatically instantiated. Between layer transitions are constrained by geometry and proximity. Motion features are used as observations. Observation likelihoods $p_{\text{in}}(z_n)$ and $p_{\text{in}}(z_n)$ are modelled as histograms and learned from data using a separate training sequence. We set $\phi()$ the same as for our indoor sequence. 160 MCMC samples are drawn at each frame with 25% burn-in samples. The birth moves are constrained to occur at receptive fields adjacent to the driver or passenger side doors.

We choose a subset of 350 frames from the "MINI-Cooper" sequence with two persons walking between the vehicles to run our proposed approach. Probabilities for birth, death, and update moves was set to $[0.0001, 0.0001, 0.9999]$ as within this sequence the number of persons to track is constant. In Fig. 10(a) frame 1171, we initialized the tracker to track two pedestrians at locations indicated by green arrows. A subset of the set of receptive fields that overlap with pedestrians in the first frame are chosen as initial particles. In frame 1221, despite severe occlusion the particle set concentrates around the true location. In frame 1386 uncertainty in depth for one pedestrian causes particles (dark blue) to diffuse a bit. In frame 1514, particles are correctly concentrated on both sides of

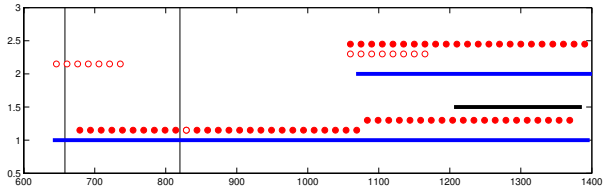


Figure 8. Tracking result time-lines for "far-field" sequence, Fig. 10(b). Please refer to Fig. 6 for plot annotation.

the MINI Cooper. We do not know of any other system that can correctly place two closely-spaced targets on both sides of an occluding layer in such a setting.

For the "far-field" sequence, we use 350 frames with interesting activity shown in Fig. 10(b) (and time-line shown in Fig. 8). A vehicle (indicated by the orange arrow) passes as pedestrians (green arrows) get out of parked cars. Birth, death, and update move probabilities was set to $[0.001, 0.0005, 0.9985]$ so the system created and deleted tracks on its own. A person in the farthest car remains seated after opening the driver-side door, his track hence latches on to the moving car and stops as the car moves out of the receptive field area. The person exits his car around frame 1070, his track is picked up and continues till end of the sequence. The driver of the mini-van in the mid-field walks around his vehicle as the car passes. His track is not lost because the receptive fields activated by the car do not have transitions from the person's location. Instead, a new short track is created. As in the case of the "MINI-Cooper" sequence, inferred hidden states reveal proximity to driver-side door and the passenger-side door. Takala's tracker shown with the thick black line only picks up one person after he moves closer to the camera.

Our C++ code runs at $4fps$ for the laboratory video and at near video frame rate for the outdoor sequences. These results, while not extensive, nonetheless demonstrate the promise of the proposed framework. The layers of graphical models provide a much richer source of information than image-plane bounding boxes, leading to an observation model that can explicitly handle occlusions. As a result, the proposed approach can track where most competing trackers would clearly fail.

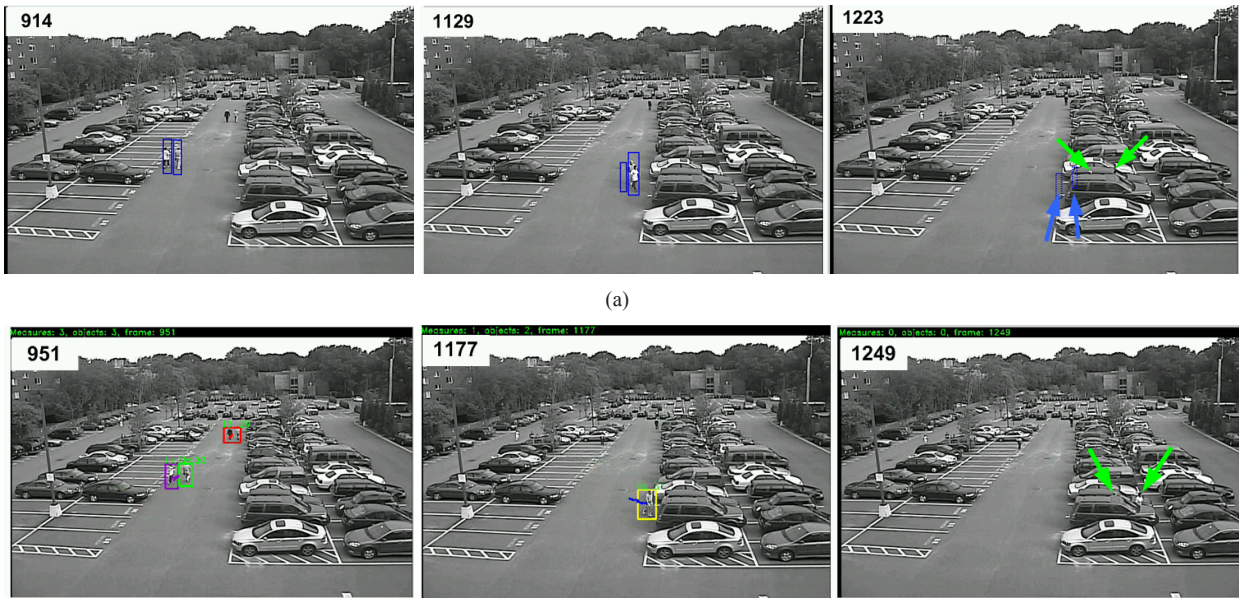


Figure 9. Tracking in continuous state space results on "MINI-Cooper" sequence using (a)Smith's [7] approach (b)Takala's [8] approach. Both trackers successfully track two pedestrians till they approach the SUV. Around this time, the two people walk on either side of the SUV with only their heads being visible causing the trackers to loose track.

6. Conclusions

We propose a graphical model for representing scene structure with static occluders employing pre-computed templates for motion around objects. Each template is represented by a graphical model. A global scene model is obtained by layering these graphical models and adding interactions between them. The richer encoding of the scene provides a layer association for the pedestrian tracks.

The full potential of our approach is perhaps best shown in the "far-field" sequence where there are variable numbers of people who open car doors and emerge at different times, (just as other vehicles pass by), resolution is poor, and severe occlusions by parked vehicles leave very few pixels for tracking. And in the MINI Cooper sequence, only foreheads remain visible, yet our method maintains track, where other blob-based trackers would fail.

Directions for future research include automatic learning of static occluding layers, e.g., via [9, 10]. The method also should be easily extended to a PTZ camera (using mosaic-building methods like [3]), since occlusion relations would be invariant for a fixed camera center. And as mentioned in the introduction, the approach is amenable to storing a database of precomputed object layer templates; this idea remains for future development.

7. Acknowledgments

We thank Valtteri Takala for sharing LBPH code and for applying [8] to our data. We thank Charles River Analytics, Inc. for the parking lot videos.

References

[1] M. Heikkila and M. Pietikainen. A texture-based method for modeling the background and detecting moving objects. *PAMI*, Vol. 28:pp. 657–662, 2006.

[2] M. Isard and J. MacCormick. Bramble: A bayesian multiple-blob tracker. In *ICCV*, 2001.

[3] J. Kang, I. Cohen, and G. Medioni. Continuous tracking within and across camera streams. In *CVPR*, 2003.

[4] Z. Khan, T. Balch, and F. Dellaert. An MCMC-based particle filter for tracking multiple interacting targets. In *ECCV*, 2004.

[5] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *ECCV*, 2002.

[6] J. Renno, D. Greenhill, J. Orwell, and G. Jones. Occlusion analysis: Learning and utilising depth maps in object tracking. *'Image and Vision Computing*, 2007.

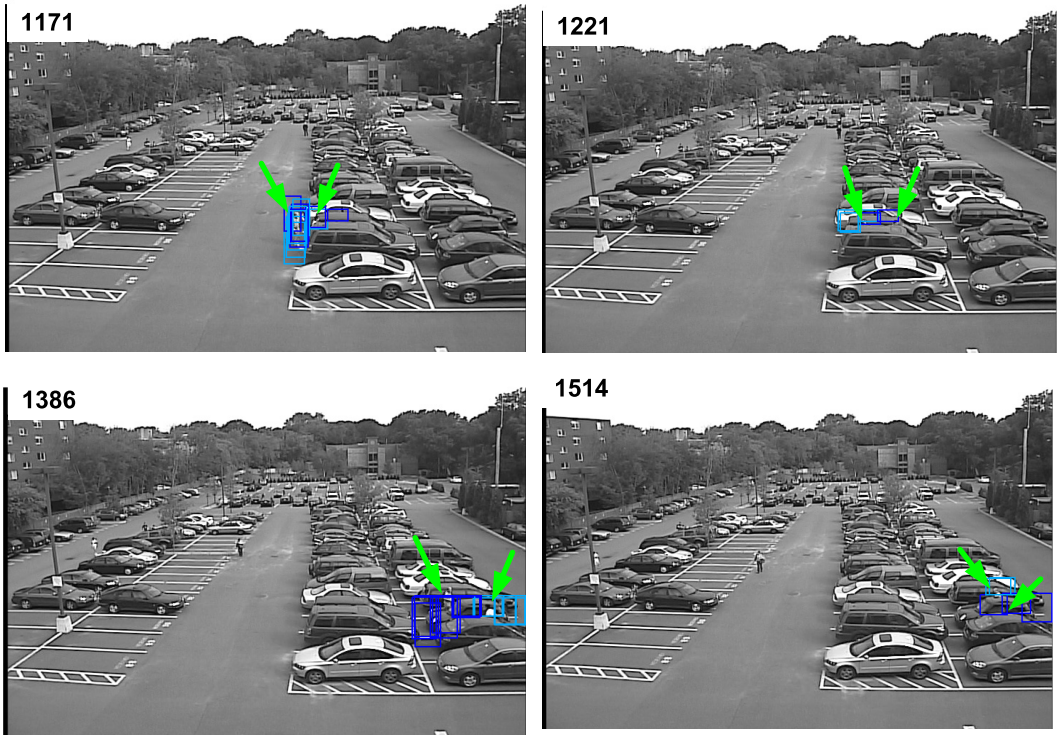
[7] K. Smith, D. Gatica-Perez, and J.-M. Odobez. Using particles to track varying numbers of interacting people. In *CVPR*, 2005.

[8] V. Takala and M. Pietikainen. Multi-object tracking using color, texture and motion. In *CVPR*, 2007.

[9] J. Winn and J. Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *CVPR*, 2006.

[10] B. Wu and R. Nevatia. Simultaneous object detection and segmentation by boosting local shape feature based classifier. In *CVPR*, 2007.

[11] Y. Zhou and H. Tao. A background layer model for object tracking through occlusion. In *ICCV*, 2003.



(a) MINI-Cooper sequence: the tracker is initialized with two pedestrians left of the SUV. The proposed approach successfully tracks and associates with the correct layer as the two persons walk on either side of the SUV and enter the MINI-Cooper.

(b) Far-field sequence corresponding to time-line in Fig. 8. A person at the back (green arrow) opening the car door is picked up but he remains seated causing the track to be terminated. His track is reacquired in the middle of the sequence with a correct layer assignment. A second person exits the mini-van in left-center of the lot and walks around the vehicle, his track is not lost as a car passes-by due to the strong person dynamics constraint. The receptive field layer for the second pedestrian is accurate over the whole sequence.

Figure 10. Tracking results for two parking lot sequences showing samples drawn from the multi-person posterior.