Crowd Detection with a Multiview Sampler

Weina Ge and Robert T. Collins

The Pennsylvania State University, University Park, PA 16802, USA

Abstract. We present a Bayesian approach for simultaneously estimating the number of people in a crowd and their spatial locations by sampling from a posterior distribution over crowd configurations. Although this framework can be naturally extended from single to multiview detection, we show that the naive extension leads to an inefficient sampler that is easily trapped in local modes. We therefore develop a set of novel proposals that leverage multiview geometry to propose global moves that jump more efficiently between modes of the posterior distribution. We also develop a statistical model of crowd configurations that can handle dependencies among people and while not requiring discretization of their spatial locations. We quantitatively evaluate our algorithm on a publicly available benchmark dataset with different crowd densities and environmental conditions, and show that our approach outperforms other state-of-the-art methods for detecting and counting people in crowds.

Keywords: Pedestrian detection; RJMCMC; Multiview geometry.

1 Introduction

Crowd detection is challenging due to scene clutter and occlusions among individuals. Despite advances in detecting and tracking people in crowds, monocular techniques are limited by ambiguities caused by insufficient information from a single view. Multiview approaches, on the other hand, can resolve ambiguities using complementary information from different views of the same scene. For example, two people totally overlapping in one view might be well separated in another view, making detection easier.

We present a probabilistic approach to estimate the *crowd configuration*, i.e. number of individuals in the scene and their spatial locations, regardless if people are visible in one view or multiple views. Our approach uses a stochastic process, specifically a Gibbs point process, to model the generation of multiview images of random crowd configurations. The optimal crowd configuration is estimated by sampling a posterior distribution to find the MAP estimate for which this generative model best fits the image observations. An overview of our approach is illustrated in Figure 1.

Our approach is motivated by the success of previous generative models for people detection [1,2,3]. Due to the great flexibility offered by sampling-based inference methods, our crowd model can accommodate inter-person dependencies that otherwise would be intractable to infer because of their inherent combinatorics. Efficient sampling strategies are the key to performance in practice.

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part V, LNCS 6315, pp. 324–337, 2010.

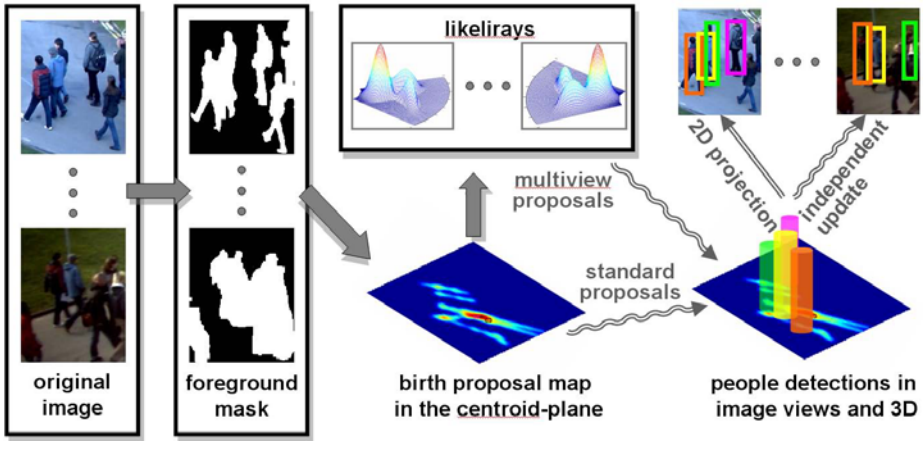

Fig. 1. Our proposed method tests hypothesized crowd configurations in 3D space against multiview observations (foreground masks) within a sampling framework

Although various data-driven proposals have been designed in the single view context to guide hypothesis generation [2,3], to our knowledge we are the first to explore multiview geometry constraints for efficient sampling.

Summary of Contributions

1. We extend generative sampling-based methods from single view to multiview, providing a unified framework for crowd analysis that successfully estimates 3D configurations in monocular and multiview input.
2. We introduce novel proposals based on multiview geometric constraints, yielding a sampler that can effectively explore a multi-modal posterior distribution to estimate 3D configurations despite occlusion and depth ambiguity.
3. Our global optimization does not require discretization of location and respects modeled spatial dependencies among people, resulting in better detection and localization accuracy than current state-of-the-art.

2 Related Work

Among monocular approaches for pedestrian detection [4,5,6,7,8,9], classifier-based methods are very popular [7,8,9] and sampling-based methods have also been shown effective for crowd detection [2,3,10] as well as generic object detection[11,12]. Within the sampling framework, various efficient, data-driven sampling strategies have been proposed. For example, Zhao and Nevatia [2] use a head detector to guide location estimates and Ge and Collins [3] learn sequence-specific shape templates to provide a better fit to foreground blobs. We extend the sampling framework to a unified approach that can detect people visible in a single view or in multiple views.

Previous multiview detection methods differ not only in image features and algorithms, but also camera layout. We confine our discussion to multiple cameras with overlapping viewpoints, for we are primarily interested in resolving ambiguities due to occlusion. Mittal and Davis [13] match color regions from all pairs of camera views to generate a ground plane occupancy map by kernel density estimation. In Khan et.al. [14], foreground likelihood maps from individual views are fused in a weighted average fashion based on field-of-view constraints. Tyagi et.al. [15] develop a kernel-based 3D tracker that constructs and clusters 3D point clouds to improve tracking performance.

Among related approaches that estimate ground plane occupancy [1,16,17,18], our work bears the closest resemblance to [1] in that we both take a generative approach. However, they discretize the ground plane into a grid of cells, and approximate the true joint occupancy probability of the grid as a product of marginal probabilities of individual cells, under the assumption that people move independently on the ground plane. Although our problem and framework are similar, we use a sampling-based inference technique that allows us to use a more flexible crowd model. Our model relaxes the independence assumption among people and does not require discretization of spatial location nor a fixed size for each person. We show in our results that these improvements lead to better detection and localization accuracy as well as greater robustness to errors in foreground estimation and camera calibration.

Our efficient sampling algorithm is inspired by previous work that seeks to improve the mixing rate of a sampler by encouraging traversal between different modes of the target distribution [19,20,21]. Dellaert et.al. [19] developed a chain flipping algorithm to generate samples of feasible solutions for weighted bipartite matching. Other methods such as the mode-hopping sampler [21] use knowledge about the topography of the target distribution to speed up sampling. Although inspiring, these methods are not directly applicable to our scenario because we are searching a large configuration space with variable dimension. More relevant is the data-driven MCMC framework [22] that uses various data-driven proposals such as edge detection and clustering to speed up Markov chain sampling for image segmentation.

3 A Gibbs Point Process for Crowd Detection

In this section we present a Bayesian statistical crowd model that accommodates inter-person dependence, together with a baseline sampling algorithm that directly extends a single view detection approach to perform multiview inference. We discuss the limitations of this baseline algorithm in Section 4 where we present the motivation and strategies of our novel multiview proposals. Experimental results on a public benchmark dataset are presented in Section 5.

3.1 Modeling

Our goal is to estimate a 3D *crowd configuration* based on image observations from a surrounding set of fixed cameras. A crowd configuration is an unordered

set of targets $\mathbf{o}^n = \{o_1, \ldots, o_n\}$, $i = 1, \ldots, n$, $n \geq 0$. Each target represents a person moving on a flat ground plane and is parameterized by an upright cylinder $o = (c, r, h)$, where $c \in W$ is a spatial coordinate in the centroid-plane, a plane that is half the height of an average person above the ground, W is a compact subset of $\mathbb{R}^2$ equipped with volume measure ν, and $[r, h]$ specifies the width (radius) and height of a person.

The configuration space is denoted as $\Omega_N = \{\emptyset, \cup_{i=1}^{N} \mathbf{o}^i\}$, which is a union of subspaces with varying dimensions, including the empty set and up to N people distributed over W. We model random configurations by a spatial point process, specifically, the Gibbs point process [23]. Let $\mu(\cdot)$ be the distribution of a homogenous Poisson process of unit intensity, which is analogous to the Lebesgue measure on $\mathbb{R}^d$. The density of the Gibbs point process can be defined with respect to this reference Poisson process. Formally,

$$p(\mathbf{o}) = \frac{f(\mathbf{o})}{\int_{\Omega} f(\mathbf{o}) d\mu(\mathbf{o})}, \tag{1}$$

where the mapping $f(\mathbf{o}) : \Omega \rightarrow [0, \infty)$ is an unnormalized density having the Gibbs form $f(\mathbf{o}) = \exp\{-U(\mathbf{o})\}$.

The Gibbs process is very flexible for modeling prior knowledge about object configurations. It often includes a unary data term to model object attributes and higher-order interaction terms to model inter-object relationships. Our model incorporates two types of inter-person dependency. The first one is an avoidance strategy motivated by studies in social science showing that people keep a 'comfort zone' around themselves. We incorporate this dependency by a Strauss Model [23], which defines a pairwise potential interaction as

$$\phi(o_i, o_j) = \begin{cases} \eta & \| c_i - c_j \| \leq r \\ 0 & \| c_i - c_j \| > r \end{cases}, \tag{2}$$

where r is a parameter that controls the size of the comfort zone and η is set to some large constant number.

The second modeled dependency is based on the principle of non-accidental alignment. It penalizes configurations where people line up perfectly along a viewing ray to claim the same foreground region. This is not a hard constraint: certainly one person can be occluded by another in any view. However, each person is unlikely to be occluded in every view. In general, we seek to penalize configurations that require a large number of occlusions to explain the data. Unfortunately, explicit occlusion analysis involves a combinatorial number of interacting subsets. To keep the energy function linear in the number of targets, we measure the degree of alignment in 3D by the amount of overlap among projected rectangles in each image view. Formally, a 'label' image is computed by pixel-wise disjunction as $\mathbf{S}^v(\mathbf{o}) = \cup_i \mathcal{H}^v(o_i)$, where $\mathcal{H}^v$ is projection function associated with camera v that maps a 3D person to a binary image that is zero everywhere except for a rectangular area bounding the projected person and $v \in [1, V]$ where V is the number of camera views. Pixels in the label image covered by at least one projected rectangle are labeled as foreground. For simplicity, we

use $\mathbf{S}^v$ as a shorthand for $\mathbf{S}^v(\mathbf{o})$. For each object o_i, $D_i^v = |\mathcal{H}^v(o_i) \cap \mathbf{S}^v(\mathbf{o} \backslash o_i)|$ measures the amount of overlap between one target's projection and the rest of the targets in that view by counting the number of foreground pixels in the intersection image. We define the overlap cost for o_i as

$$D_i = \begin{cases} D_i^v & o_i \text{ only visible in } v \\ \min\limits_v D_i^v & \text{otherwise} \end{cases}.$$

This way, overlap in some views will not be penalized as long as the target is clearly visible in other views. We encode prior knowledge about a general crowd configuration in the total energy of a Gibbs process

$$U(\mathbf{o}) = \sum_{i,j} \phi(o_i, o_j) + \sum_i D_i + \gamma N, \tag{3}$$

where $N = |\mathbf{o}|$ is the number of estimated people in 3D. The last term penalizes spurious detections with a constant weight γ.

Under this probabilistic framework, the problem of crowd detection is solved by finding the configuration that best explains the image observations (foreground masks) from different views. Denote the binary foreground mask in view v by $\mathbf{Z}^v = \{Z_i^v\}$, $Z_i^v \in \{0, 1\}$, $i = 1, \ldots, m_v$, where m_v is the number of pixels in the image observed from view v. A likelihood function $\mathcal{L}$ is defined to measure the probability of a configuration given the foreground masks by comparing two sets of binary images, the mask images $\mathbf{Z}$ and label images $\mathbf{S}$,

$$\mathcal{L}(\mathbf{o}; \mathbf{Z}) = \mathcal{L}(\mathbf{S}; \mathbf{Z}) = \exp\{-G(\mathbf{o})\}, \tag{4}$$

$$G(\mathbf{o}) = \textstyle\sum_{v=1}^{V} \sum_{i=1}^{m_v} I_1(S_i^v, Z_i^v) + \beta \sum_{j=1}^{N} I_2(o_j), \tag{5}$$

$$I_1(S_i^v, Z_i^v) = \begin{cases} 1 & S_i^v \neq Z_i^v \\ 0 & \text{o.w.} \end{cases}, \quad I_2(o_j) = \begin{cases} 1 & \exists v, \text{ s.t. } \frac{|\mathcal{H}^v(o_i) \cap \mathbf{Z}^v|}{|\mathcal{H}^v(o_i)|} < 0.1 \\ 0 & \text{o.w.} \end{cases}, \tag{6}$$

This likelihood function contains two terms: I_1 penalizes discrepancies between hypothesized person detections and the image observations, and I_2 imposes an extra penalty on 'ghosts' – detections that cover mostly background pixels. β is set to some large constant number.

Combining the prior (Eqn. 1) and the likelihood function (Eqn. 4), we define the optimal crowd configuration as the MAP estimator

$$\mathbf{o}^* = \arg\max_{o \in \Omega}(P(\mathbf{o}|\mathbf{Z})) = \arg\max_{o \in \Omega}\Big(\frac{e^{-\left(U(\mathbf{o}) + G(\mathbf{o})\right)}}{C(\Omega)}\Big). \tag{7}$$

Optimizing the above posterior directly is intractable because the normalizing constant from the Gibbs prior, $C(\Omega) = \int_\Omega f(\mathbf{o}) d\mu(\mathbf{o})$, involves all possible configurations in the combinatorial configuration space Ω. Moreover, pairwise potentials in our crowd model make the inference harder than what can be handled by approximation methods such as [1,18,24].

3.2 Inference

We use reversible jump Markov Chain Monte Carlo (RJMCMC) to battle the intractable normalizing constant in Eq. 7. MCMC is designed to generate samples from complicated target distributions, such as our posterior distribution, by constructing a Markov chain with the desired target distribution as its equilibrium distribution. RJMCMC [25] extends the classic algorithm to deal with variable dimension models. It suits the crowd analysis problem well because the number of people is not known apriori, and thus also needs to be estimated.

The RJMCMC sampler explores the configuration space by proposing perturbations to a current configuration. The general sampling framework is reviewed in the supplemental material[1]. The design of good proposal distributions is the most challenging part of the sampling algorithm. Proposals that only allow local perturbations may become trapped in local modes, leaving large portions of the solution space unexplored, whereas global adjustments have less chance to be accepted unless the target distribution is very smooth or tempered to be so. To achieve a good balance of both local and global proposals, we use proposals from a mixture of both types: $Q(;) = \sum_{c=1}^{C} p_c Q_c(;)$, where $\sum_c p_c = 1, \int Q_c(\mathbf{o}'; \mathbf{o})\mu(d\mathbf{o}') = 1$, and C is the number of different proposal moves. Below we describe a baseline multiview sampler directly extended from local birth, death, and update proposals commonly used in single view samplers [2,3].

Birth/Death proposal. A birth proposal adds a 3D person to the current configuration, i.e. $\mathbf{o}' = \mathbf{o} \cup o_b$. A simple birth strategy might place a person uniformly at random (u.a.r.) in the bounded region W. A death proposal removes a person from the current configuration so that $\mathbf{o}' = \mathbf{o} \setminus o_d$, e.g. choosing o_d u.a.r. from $\mathbf{o}$. Both proposals involve a dimension change from $|\mathbf{o}|$ to $|\mathbf{o}'|$. Instead of blindly adding a person, we use a more informative data-driven proposal [22]. We sample o_b's location according to the birth probability $P_b \sim \frac{P_b(l)}{\sum_{l \in \widetilde{W}} P_b(l)}$, where $P_b(l) = \frac{1}{V} \sum_v \frac{|\mathcal{H}^v(l) \cap Z^v|}{|\mathcal{H}^v(l)|}$ is the fused occupancy likelihood of a particular location l, computed as the sum of the percentage of foreground pixels within its projected rectangles in all views, and $\widetilde{W}$ is a discretization of the bounded region of interest in the centroid-plane W. Our final detection results are not restricted by this discretization because localization is adjusted by other proposals of the sampler.

Update Proposal. The update proposal preserves the dimension of the current configuration but perturbs its member's attributes (location and size) to generate a new configuration. We use a random walk proposal that selects a person o_u u.a.r. from $\mathbf{o}$, and either proposes a new spatial placement by sampling from a truncated normal distribution $\mathcal{N}(c'|c_u, \sigma)$ centered at the current location c_u, or proposes a new size by sampling from a truncated normal centered at the size of an average person, $h = 1.7m$ and $r = 0.4m$.

[1] http://vision.cse.psu.edu/projects/multiviewmcmc/multiviewmcmc.html

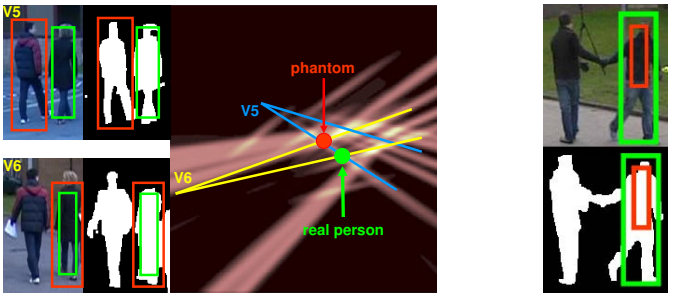

Fig. 2. Common pitfalls in multiview crowd detection. **Left**: the phantom phenomenon. A 3D phantom location (red circle) explains foreground pixels in different views that actually belong to the projections of two different real people (red boxes). **Right:** depth ambiguity for people visible in a single view can result in explanation of a single foreground region by alternative detections of different sizes.

4 Multiview Proposals

The local proposals presented in the previous section yield satisfactory results when people are well-separated in multiple views. However, when a person is visible only in one view, the inherent depth ambiguity coupled with noisy foreground blobs leads to a significant performance drop, which has also been reported in previous work [18,24]. Moreover, as occlusion becomes more frequent, we have observed that the naive sampler often gets stuck in local modes because of the 'phantom' phenomenon. Phantoms are accidental intersections of viewing rays at locations that are not occupied by any real person. Phantom hypotheses attempt to explain foreground regions across multiple views that actually are projections of different people in 3D. As shown in Figure 2, when a phantom gets accepted in the current configuration, later proposals for the real person are less likely to get accepted because the phantom already explains a large portion of their foreground pixels, thus the new birth proposal will suffer a high overlap penalty. Local random walk updates are also unlikely to escape from this local maximum. Although increasing the step size of a random walk can alleviate the problem to some extent, such blind exploration wastes time visiting mostly low probability regions, leading to an inefficient sampler.

Inspired by long range mode-hopping MCMC proposals [19,20,21], we exploit geometric constraints to design proposals that allow global changes that more effectively explore the configuration space. The motivation behind using geometric constraints is that multiview geometry is consistent across views whereas image-based appearance constraints (e.g. head detection for birth [2]) may conflict with each other in different views.

Our multiview proposals are based on occupancy likelihood rays, or *likelirays* for short. Recall that in our data-driven birth proposal, we have computed a centroid-plane occupancy map by fusing foreground masks from all the views in 3D. Likelirays are essentially polar coordinate transformations of the

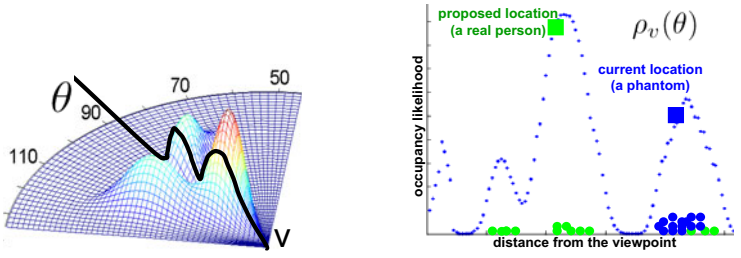

Fig. 3. Left: Likelirays from one viewpoint v, indexed by angle θ, define a set of 1D distributions over potential person and phantom locations at different depths along viewing rays in the centroid-plane. **Right:** Mode-hopping by sampling from a likeliray $\rho_v(\theta)$. Green dots are samples from the depth move, which visits all significant modes whereas the blue samples from a local random walk proposal stays in a single mode.

centroid-plane occupancy map with respect to each camera view v, indexed by angle θ, i.e. $\rho_v(\theta)$. Different modes along each likeliray correspond to potential real and phantom locations of people at different depths. The likeliray representation gives us a convenient way to generate proposals with respect to a single camera view while taking into account fused information from all other camera views. We now present two such multiview proposals.

Depth Move Proposal. A depth move first randomly selects a person o_m and a camera view v from the list of views where o_m is visible. Let θ denote the angle of the polar coordinate of o_m. A new 3D location is sampled with probability proportional to the 1D likeliray distribution $\rho_v(\theta)$. Figure 3 shows that samples from depth moves are able to visit different modes whereas samples from local random walk proposals only cluster around the current location. The depth proposal is a powerful and versatile mechanisim to handle the problems shown in Figure 2. It can switch between a phantom and a real person hypothesis and also can produce the effect of a large scale change of a single person by "sliding" them in depth along a viewing ray, which is useful for correctly detecting people visible only in a single view. Unlike random walk with large step size, a depth move preserves some of the already covered foreground pixels. Depth moves therefore tend not to cause large decreases in likelihood, so are more likely to be accepted.

Merge/Split Proposal. When people are only visible in a single view and the viewpoint is not very elevated, a large foreground region may become covered by fragmented detections corresponding to pedestrian hypotheses scattered within a small range of viewing angles at different distances from the camera may be hypothesized to cover parts of a large foreground region (Figure 4). These fragments create local modes that prevent explaining the entire region correctly as one single person.

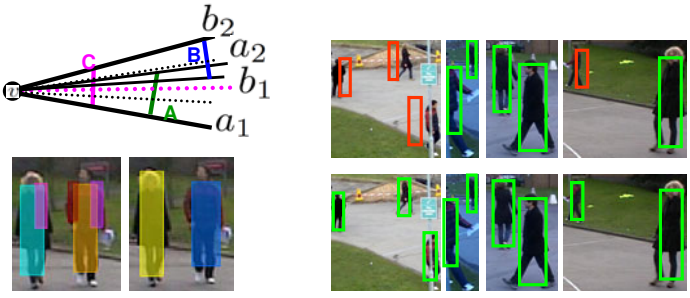

Fig. 4. Left: Merge Proposal. The top panel shows how the 3D merge proposal yields a new hypothesis C that minimally covers both projections A and B in view v. The bottom shows that the final results (right) correctly recover from fragmented hypotheses (left). **Right:** Independent Update Proposal. The top panel shows localization error (marked in red) in four views due to camera calibration and synchronization errors. The bottom shows improved results using the independent update proposal.

We design a 3D merge/split move to ease the switch between the following two hypotheses: multiple distant people versus a single, closer person. Let two people o_a and o_b both be visible from a particular viewpoint, with polar coordinates (θ_a, r_a) and (θ_b, r_b), $\theta \in (0, \pi)$. As illustrated in Figure 4, their angular extents are $[a_1, a_2]$ and $[b_1, b_2]$. A new merged person o_c can be hypothesized from o_a and o_b in two ways: 1) when one of the angular extents completely falls within the other, we randomly move the person with the larger angular extent closer to the camera and delete the other; 2) otherwise, without loss of generality, assume $a_1 < b_1$ and $a_2 < b_2$, which includes the case of partial occlusion as well as complete separation of the two. We create a new person in 3D whose image projection minimally covers the projections of both merge candidates, thus having an angular extent $[a_1, b_2]$. The corresponding polar coordinates (θ_c, r_c) of o_c can be computed as $\theta_c = \frac{a_1+b_2}{2}$, $r_c = \frac{0.5w}{\tan(0.5(b_2-a_1))}$, where w is the width of an average sized person.

A 3D merge move randomly chooses a view v in which to propose a merge. Denoting all visible people in v as $\mathbf{o}_v$, a person o_a is chosen u.a.r. from $\mathbf{o}_v$. For each other person o_i, $i \neq a$, let e_i be the angular extent of the candidate blob that would result from merging o_a and o_i. We use these extents to define a probability distribution over candidates i as $p_i = \frac{\tilde{e}_i}{\sum_j \tilde{e}_j}$, where $\tilde{e}_i = \frac{\min_j e_j}{e_i}$ favors merging two people with large angular overlap. A candidate person is proposed for merging with o_a by sampling from this distribution. If a newly merged person is accepted, we store their components in a merge list. The reverse proposal is a 3D split that randomly selects a person from the merge list and splits them back into their stored original component detections.

Independent Update Proposal. So far, the four types of presented proposals, birth/death, update, depth move, and merge/split, all hypothesize new person locations/sizes in 3D and the corresponding projections in image views are determined by the camera calibration information. To accommodate noisy input,

e.g. errors in calibration, synchronization, or foreground estimation, we add an independent update proposal that can perturb the 2D projection rectangle in each image plane independently (demonstrated in Figure 4). The independent update move works by randomly choosing a person o_i and a camera view v from the list of views where o_i is visible. With equal probability, either the size or the location of the projection box in view v is updated by sampling from a truncated 2D normal distribution centered at the nominal image location and size determined by the calibration matrices.

5 Experiments

We evaluate our algorithm on the PETS2009 dataset [26], a challenging benchmark dataset for multiview crowd image analysis containing outdoor sequences with varying crowd densities and activities. We tested on two tasks: crowd detection in a sparse crowd (sequence S2L1-1234) and crowd counting in a dense crowd (sequence S1L1-1357). We generated foreground masks using an adaptive background subtraction algorithm similar to Zivkovic's method [27], and camera calibration information provided with each dataset was used to generate the birth proposal map P_b as the average back-projection of foreground masks from all views, as described in Section 3.2. Sample detection results are shown in Figure 6. Our proposed method obtains superior results over other state-of-the-art crowd detection methods, as will be shown through quantitative evaluation below.

Sparse sequence S2L1: We used four camera views, including one elevated, far field view (called View 1) and three low-elevation near field views with frequent, severe occlusions (Views 5, 6, and 8). We compared our detection results against the ASEF method, which is a detection method using convolution of learned average of synthetic exact filters [5], and the POM+LP method, which is a multi-target detection and tracking algorithm based on a probabilistic occupancy map and linear programming [24]. We chose these two methods because they are the current top-performers as reported in Winter-PETS2009 [26]. We also compared against the Cascade [8] and Part-based [9] person detectors, trained according to [5]. We performed ground-truth annotation of the sequence and evaluated each algorithm based on the standard MODA and MODP metrics (details are included in the supplemental material[1]). MODP measures localization quality of the correct detections and MODA measures detection accuracy taking into account false negatives/positives. For both metrics, larger values are better. A detection is counted as correct if the overlap ratio between the annotated box and the detection box is greater than some threshold τ. We systematically vary this threshold and compute the evaluation metrics at each threshold. Correct detections and false positives/negatives are determined by solving an assignment problem between the annotations and the detection output.

Figure 5(A) shows averaged MODP and MODA scores across four views for our method and POM+LP, and over the detections from View 1 for the three classifier-based detectors (those are monocular methods that only have results

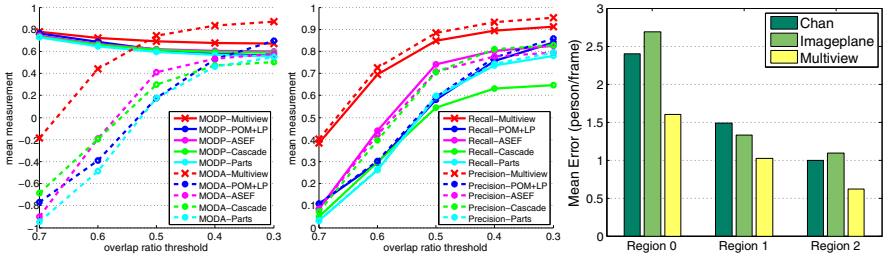

Fig. 5. Evaluation results on S2L1 and S1L1. For S2L1, our algorithm (red curves) consistently outperforms other methods in terms of MODA&MODP (**A**) and Precision&Recall metrics (**B**) at different overlap threshold levels without using temporal or appearance information. For S1L1 (**C**), we achieve lower count errors in all three target regions than current state-of-the-art methods.

reported for View 1). Our multiview MCMC method consistently outperforms others (higher detection accuracy) at all overlap threshold levels. Additionally, the prominent performance gap at the tighter end of the threshold levels (larger τ) indicates that our method has better localization quality than other methods. It is interesting to note that our method is the top performer even though we do not use temporal consistency constraints across frames or discriminative object appearance information. Our improved accuracy is due to use of a more flexible generative model, made possible by the sampling-based inference, and our novel multiview proposals that allow more efficient global exploration of the posterior distribution. Since we are free from restrictions of discrete ground-plane grids, fixed 3D person size, and independence among people, we achieve better 3D localization than POM+LP, even with noisy input (Figure 6).

As the overlap threshold decreases, we see (as expected) an increase in MODA and decrease in MODP, since more misaligned detections become classified as correct. However, our MODP curve has a slower decreasing rate than others, which again confirms that we achieve better localization accuracy. Figure 5(B) shows results from a similar comparison but using precision/recall metrics. Our method has higher recall and precision than other methods.

In Table 1, we compare our multiview MCMC method to the naive baseline MCMC approach, introduced in Section 3.2, which does not use the new multiview proposals. The new multiview method outperforms the baseline approach in all cases. In the same table, we also show that our method works well with monocular sequences. For this experiment, we only use input observations from a single view. As opposed to the significant performance drop of the POM method reported in [24] in this situation, our single view detection results do not vary dramatically from the multiview results, and continue to outperform the other methods. These experiments indicate that our multiview proposals are effective at dealing with depth ambiguity, even along viewing rays from a single camera.

Table 1. MODP (1st column) and MODA (2nd column) in each view of S2L1 at an overlap threshold τ of 0.5. Scores in bold indicate the top-ranked algorithm with respect to score metric and view. The first three rows are variants of our sampling-based approach and the bottom four are other state-of-the-art methods.

Method	View 1		View 5		View 6		View 8	
Multiview	0.6805	**0.7532**	**0.6872**	**0.6998**	**0.6953**	**0.8162**	**0.7004**	**0.6941**
Baseline	0.6791	0.6988	0.6872	0.5660	0.6936	0.6967	0.6967	0.5702
Singleview	**0.6863**	0.7052	0.6751	0.6415	0.6855	0.5333	0.6924	0.5357
POM+LP	0.5806	-0.1037	0.6071	0.2630	0.6467	0.3354	0.6344	0.2188
ASEF	0.6212	0.4116	-					
Cascade	0.6150	0.3000	-					
Parts	0.5927	0.1759	-					

Fig. 6. Sample detection results for S1L1 (top) and S2L1 (middle), overlaid on the original images and foreground masks. The bottom row shows sensitivity of our method to varying levels of noise in the foreground mask.

Dense sequence S1L1: The S1L1 sequence is captured from more elevated camera viewpoints, but with a higher crowd density and more lighting changes due to intermittent cloud cover. We annotated ground-truth person counts in all three regions specified by the PETS evaluation for View 1, shown in Figure 6,

and detect people using two camera views. The average count error for each region over the whole sequence is reported in Figure 5(C). Our error rate is less than 2 people per frame, better than the already remarkable results from Chan using holistic properties [28], which are the best results reported so far. We also compared against a 2D MCMC implementation [10] that performs birth, death and update proposals within the 2D image plane of View 1.

In Figure 6 we show sensitivity of this approach to errors in foreground estimation. This is an indoor sequence of 4 people walking [1]. On the left we see that our approach is tolerant of typical levels of foreground noise. However, as shown on the right, large areas of the image incorrectly labeled as foreground (due, for example, to failures of background subtraction to handle rapid lighting changes), can lead to false positive detections. However, our framework can be easily adapted to input data other than foreground masks, such as motion information or pedestrian classifier score maps.

6 Conclusion

We extend monocular, sampling-based crowd detection methods to perform multiview detection to accurately localize people in 3D given single or multiview images. Our results on a challenging benchmark dataset for crowd analysis demonstrate the advantage of our approach compared to other state-of-the-art methods. We have designed novel proposals that leverage multiview geometric constraints to effectively explore a combinatorial configuration space with varying dimension (numbers of people) while solving the problem of phantoms in multiview sequences and depth ambiguity in monocular sequences. Our sampling-based inference framework yields great flexibility in defining generative models that enable accurate localization of individuals in crowds despite occlusions, noisy foreground masks, and errors in camera calibration and synchronization.

Acknowledgments. We thank other PETS09 participants for sharing their results. This work was partially funded by NSF IIS-0729363.

References

1. Fleuret, F., Lengagne, R., Fua, P.: Fixed point probability field for complex occlusion handling. In: ICCV (2005)
2. Zhao, T., Nevatia, R.: Bayesian human segmentation in crowded situations. In: CVPR (2003)
3. Ge, W., Collins, R.T.: Marked point processes for crowd counting. In: CVPR (2009)
4. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: CVPR, pp. 878–885 (2005)
5. Bolme, D.S., Draper, B.A., Beveridge, J.R.: Average of synthetic exact filters. In: CVPR, pp. 2105–2112 (2009)
6. Tu, P., Sebastian, T., Doretto, G., Krahnstoever, N., Rittscher, J., Yu, T.: Unified crowd segmentation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 691–704. Springer, Heidelberg (2008)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: ICCV (2005)

8. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR, pp. 511–518 (2001)
9. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
10. Ge, W., Collins, R.T.: Evaluation of sampling-based pedestrian detection for crowd counting. In: Winter-PETS (2009)
11. Ortner, M., Descombes, X., Zerubia, J.: A marked point process of rectangles and segments for automatic analysis of digital elevation models. TPAMI 30, 105–119 (2008)
12. Rue, H., Hurn, M.: Bayesian object identification. Biometrika 86, 649–660 (1999)
13. Mittal, A., Davis, L.S.: M2Tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 18–33. Springer, Heidelberg (2002)
14. Khan, S.M., Shah, M.: Tracking multiple occluding people by localizing on multiple scene planes. TPAMI 31, 505–519 (2009)
15. Tyagi, A., Keck, M., Davis, J., Potamianos, G.: Kernel-Based 3D tracking. In: IEEE International Workshop on Visual Surveillance (2007)
16. Otsuka, K., Mukawa, N.: Multiview occlusion analysis for tracking densely populated objects based on 2-D visual angles. In: CVPR, vol. 1, pp. 90–97 (2004)
17. Yang, D.B., González-Baños, H.H., Guibas, L.J.: Counting people in crowds with a real-time network of simple image sensors. In: ICCV, pp. 122–129 (2003)
18. Alahi, A., Jacques, L., Boursier, Y., Vandergheynst, P.: Sparsity-driven people localization algorithm: Evaluation in crowded scenes environments. In: Winter-PETS (2009)
19. Dellaert, F., Seitz, S., Thorpe, C., Thrun, S.: EM, MCMC, and chain flipping for structure from motion with unknown correspondence. Machine Learning 50 (2003)
20. Andricioaei, I., Straub, J.E., Voter, A.F.: Smart darting Monte Carlo. The Journal of Chemical Physics 114, 6994–7000 (2001)
21. Sminchisescu, C., Welling, M., Hinton, G.: A Mode-Hopping MCMC Sampler. Technical Report CSRG-478, University of Toronto (2003)
22. Zhu, S., Zhang, R., Tu, Z.: Integrating bottom-up/top-down for object recognition by data driven Markov Chain Monte Carlo. In: CVPR, pp. 738–745 (2000)
23. van Lieshout, M.: Markov Point Processes and their Applications. Imperial College Press, London (2000)
24. Berclaz, J., Fleuret, F., Fua, P.: Multiple object tracking using flow linear programming. In: Winter-PETS (2009)
25. Green, P.: Reversible jump Markov chain Monte-Carlo computation and Bayesian model determination. Biometrika 82, 711–732 (1995)
26. Ellis, A., Shahrokni, A., Ferryman, J.M.: PETS 2009 and Winter-PETS 2009 results: A combined evaluation. In: Winter-PETS (2009)
27. Zivkovic, Z.: Improved adaptive gaussian mixture model for background subtraction. In: ICPR, vol. 2, pp. 28–31 (2004)
28. Chan, A., Morrow, M., Vasconcelos, N.: Analysis of crowded scenes using holistic properties. In: PETS (2009)