Multiple View Discriminative Appearance Modeling with IMCMC for Distributed Tracking

Santhoshkumar Sunderrajan, B.S. Manjunath
Department of Electrical and Computer Engineering
University of California, Santa Barbara
{santhosh,manj}@ece.ucsb.edu

***Abstract*—This paper proposes a distributed multi-camera tracking algorithm with interacting particle filters. A robust multi-view appearance model is obtained by sharing training samples between views. Motivated by incremental learning and [1], we create an intermediate data representation between two camera views with generative subspaces as points on a Grassmann manifold, and sample along the geodesic between training data from two views to uncover the meaningful description due to viewpoint changes. Finally, a Boosted appearance model is trained using the projected training samples on to these generative subspaces. For each object, a set of two particle filters i.e., local and global is used. The local particle filter models the object motion in the image plane. The global particle filter models the object motion in the ground plane. These particle filters are integrated into a unified Interacting Markov Chain Monte Carlo (IMCMC) framework. We show the manner in which we induce priors on scene specific information into the global particle filter to improve tracking accuracy. The proposed algorithm is validated with extensive experimentation in challenging camera network data, and compares favorably with state of the art object trackers.**

Keywords*—*Distributed Camera Network, Particle Filters, Object Tracking.

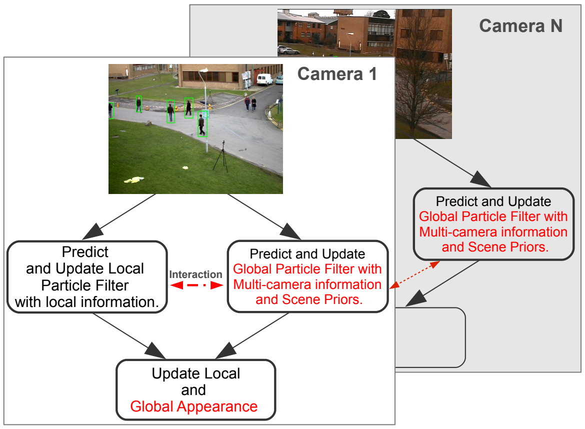

Fig. 1: **Proposed Multi-camera based tracking algorithm:** Both local and global particle filters operate in parallel. Local particle filter interacts with the global particle filter using IMCMC framework [6]. Local particle filter takes only local information available within the camera view into account. Global particle takes multi-camera information (performs fusion) and scene priors into account. Finally, a robust global appearance model is learnt by sharing training samples with the neighboring camera views. Contributions of this paper are highlighted in red.

I. Introduction

Proliferation of cheap and network enabled smart cameras has provided an enormous opportunity for large scale deployment of camera networks in real-world applications. This paper proposes a novel tracking algorithm to track pedestrians using a set of cameras with overlapping views by integrating information from different camera views. The ground plane homography is used to relate information obtained from multiple cameras to estimate an object's position in the ground plane. Further, to enhance tracking quality, scene priors such as crowd flow information from the past historical data is used. More importantly, we propose a strategy to learn a robust discriminative appearance model by sampling training examples from generative subspaces between two views and this leads to significant improvement in tracking accuracy.

Existing multi-camera tracking algorithms assume that the tracking problem in individual cameras is solved, and focus on higher level tasks such as activity recognition, event analysis and camera-hand-off. Due to background distortions, complex illumination changes, varying object shapes and packet losses in wireless communication, the performance of single camera tracking is far from the ideal. Also, the tracking algorithm is most likely to fail due to improper appearance modeling. Multiple viewpoint changes are not explicitly modeled in existing tracking algorithms. More importantly, information fusion algorithms do not take scene priors into account.

There are a number of multi-camera tracking algorithms [2], yet very little attention has been paid to distributed tracking algorithms. Existing distributed tracking algorithms [3], [4], [5] manifest as solutions to an information fusion problem and do not take prior information about the network into consideration during the fusion. Also, not much attention has been paid to robust multi-camera appearance modeling. We propose a novel strategy to train a discriminative Boosted appearance model by sharing training samples with neighboring views by taking view-shift into account. More importantly, we induce the priors on scene information such as crowd flow into a particle filter framework and demonstrate that this significantly improves the robustness of object tracking.

For a synchronized network with M cameras, let $\{\mathbf{I}^{c}_{1:t}\}$ be a set of video frames from different camera views where $c \in \{1 \dots M\}$. At time instance t, the state of the i^{th} object on the image plane of the c^{th} camera is denoted by $\mathbf{X}^{c,i}_{t} = [positionX, positionY, sizeX, sizeY]$ where $i \in \{1, \dots, O\}$ is the global object index and O is the number of objects initialized by the object detector at the first frame. $[positionX, positionY]$ is the center of the object's bounding box on the image plane and $[sizeX, sizeY]$ is the size of the object along the x and y directions. Given the set of video frames, i.e. $\{\mathbf{I}^c_{1:t}\}$ from different cameras, we infer the object state for the i^{th} object, $\{\mathbf{X}^{c,i}_t\} = [\mathbf{X}^{1,i}_t \ldots \mathbf{X}^{M,i}_t]$ by Maximum a Posteriori (MAP) formulation:

$$\arg\max_{\{\mathbf{X}^{c,i}_t\}} p(\{\mathbf{X}^{c,i}_t\}|\{\mathbf{I}^c_{1:t}\}) \tag{1}$$

where i is the object index and c is the camera index. We propose a novel strategy to train a robust discriminative multi-view appearance model $p(\mathbf{Y}^{c,i}_t|\mathbf{X}^{c,i}_t)$, by taking viewpoint changes into account. Additionally, we propose a unified Markov Chain Monte Carlo (MCMC) framework to combine local and global information into the tracker. Most importantly, we show the manner in which we induce the scene specific priors and multi-camera interaction into the global particle filter to improve the tracking accuracy. Figure 1 highlights the contributions of the proposed multi-camera tracking algorithm. Following are the main contributions of this paper:

1. A robust globally discriminative appearance modeling by taking viewpoint shift into account.
2. A unified MCMC approach to combine local and global models. Global particle filter models the multi-camera information fusion and takes scene priors into account.

II. Related Works

In [7], a distributed Kalman filtering framework is used to track the objects on the global ground plane. There have been recent efforts to perform active collaboration between cameras. [8] proposed a Bayesian algorithm for distributed multi-target tracking using multiple collaborative cameras and they used 2D location instead of 3D. They do not take higher level scene information into account while performing the fusion. Especially in [9], a distributed fusion mechanism that clusters particles obtained from multiple camera views is proposed. This method is highly vulnerable to outliers in shared particles. In contrast, our distributed tracking algorithm combines local and global motion models in a unified probabilistic framework. Also, scene priors are taken into account while updating the global motion model. There are a few single camera tracking algorithms that take scene priors into account to improve the tracking accuracy, however, these methods are not straightforward to extend to multi-camera distributed tracking scenarios [10], [11].

Recently, tracking by detection algorithms have been gaining popularity [12], [13], [14]. Existing multiple camera tracking algorithms do not discriminatively model the multi-view appearance in an online manner. Roth et al. [15] proposed a multiple instance learning based co-training strategy for multi-view appearance tracking. In comparison, we propose a novel distributed strategy to train a discriminative appearance model by taking viewpoint shift into account.

The rest of the paper is organized as follows. Section III introduces Bayesian Tracking formulation with Markov Chain and Monte Carlo. Section IV formulates the multi-camera tracking algorithm. Section IV-A proposes a novel method for multi-view appearance modeling and section IV-B discusses multi-camera information fusion by inducing scene priors. Section V demonstrates the efficacy of the proposed methodology on some challenging multi-camera datasets and finally we conclude the paper in section VI.

III. Bayesian Tracking Formulation

We explain the basics of particle filter based object tracking [11]. The goal of object tracking is to find the best object configuration $\mathbf{X}_t$ given the observations $\mathbf{Y}_{1:t}$ up to time t using the following Bayesian formulation:

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t}) \propto p(\mathbf{Y}_t|\mathbf{X}_t) \int p(\mathbf{X}_t|\mathbf{X}_{t-1}) p(\mathbf{X}_{t-1}|\mathbf{Y}_{1:t-1})\, d\mathbf{X}_{t-1}, \tag{2}$$

The optimal object configuration $\hat{\mathbf{X}}_t$ is obtained by Maximum a Posteriori (MAP) estimation:

$$\hat{\mathbf{X}}_t = \arg\max_{\mathbf{X}_t}\ p(\mathbf{X}_t|\mathbf{Y}_{1:t}) \tag{3}$$

The posterior at time "$t-1$", is approximated by a set of weighted particles in "Sequential Importance Resampling" (SIR) particle filters:

$$p(\mathbf{X}_{t-1}|\mathbf{Y}_{t-1}) \approx \{\mathbf{X}^{(p)}_{t-1}, \pi^{(p)}_{t-1}\}^P_{p=1} \tag{4}$$

where p is the particle index and P is the number of particles. The weight of the p^{th} particle is given by, $\pi^{(p)}_t = p(\mathbf{Y}_t|\mathbf{X}^{(p)}_t)$. The integral in the equation 2 can be approximated by:

$$p(\mathbf{X}_t|\mathbf{Y}_t) \approx p(\mathbf{Y}_t|\mathbf{X}_t) \sum_{p=1}^{P} \pi^{(p)}_{t-1} p(\mathbf{X}_t|\mathbf{X}^{(p)}_{t-1}) \tag{5}$$

We use MCMC sampling instead of the standard "importance resampling". Compared to "importance resampling" based particle filter, the MCMC based particle filter is very efficient [16]. The complexity varies linearly with respect to number of objects compared to exponentially varying complexity in SIR particle filters. More importantly, "importance resampling" methods suffer from particle impoverishment and degeneracy. In MCMC methods, a Markov chain is defined over the space of configurations $\mathbf{X}_t$ and the stationary distribution of the chain is equal to the posterior distribution, $p(\mathbf{X}|\mathbf{Y})$. A set of un-weighted samples i.e., $p(\mathbf{X}_t|\mathbf{Y}_t) \approx \{\mathbf{X}^{(p)}_t\}^P_{p=1}$ is used to represent the posterior in MCMC based particle filter.

IV. Distributed Multi-object Tracking in a Camera Network

For a camera network with M synchronized cameras, the primary task is to jointly track objects with active collaboration between the views. The goal is to estimate the state of objects, i.e, $\{\mathbf{X}^{c,i}_t\}$ given the set of image observations $\{\mathbf{Y}^{c,i}_{1:t}\}$ on the set of video frames $\{\mathbf{I}^c_{1:t}\}$ from different cameras and ground plane homography. We assume that no raw image data is transferred between nodes due to network bandwidth constraints and estimate the following by approximating equation 1 on camera $c = C$ for the i^{th} object as:

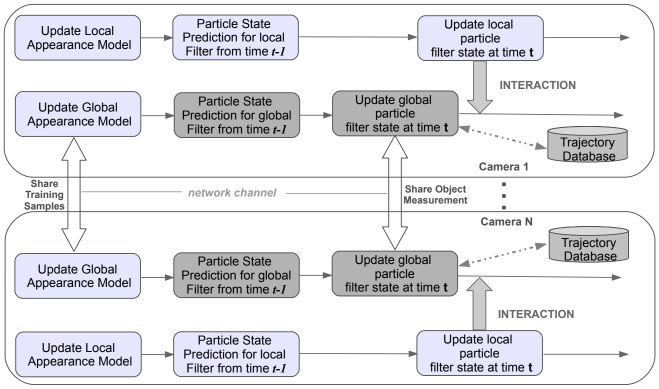
Update Local Appearance Model
Particle State Prediction for local Filter from time t-1
Update local particle filter state at time t
INTERACTION
Update Global Appearance Model
Particle State Prediction for global Filter from time t-1
Update global particle filter state at time t
Trajectory Database
Camera 1
Share Training Samples
network channel
Share Object Measurement
Camera N
Update Global Appearance Model
Particle State Prediction for global Filter from time t-1
Update global particle filter state at time t
Trajectory Database
INTERACTION
Update Local Appearance Model
Particle State Prediction for local Filter from time t-1
Update local particle filter state at time t

Fig. 2: **Interactive MCMC based tracking:** At time $t-1$, the global appearance model is trained by sharing training examples between different views. The local appearance model is trained using examples within the image view. Both local and global particle filters predict the state from time $t-1$. The global particle updates its state using the object measurement shared between different views. The local particle filter interacts with the global particle filter about the best configuration of the object.

$$\arg\max_{\{\mathbf{X}_t^{c,i}\}} p(\{\mathbf{X}_t^{c,i}\}|\mathbf{I}_{1:t}^{c=C}), c = 1\ldots M \tag{6}$$

By assuming $\mathbf{I}_t^{c=C}$ is conditionally independent of $\mathbf{I}_{1:t-1}^{c=C}$ given the estimates $\{\mathbf{X}_t^{c,i}\}$ and expanding equation 6 by Bayes rule, we have the following:

$$p(\{\mathbf{X}_t^{c,i}\}|\mathbf{I}_{1:t}^{c=C}) \propto p(\mathbf{I}_t^{c=C}|\{\mathbf{X}_t^{c,i}\})p(\{\mathbf{X}_t^{c,i}\}|\mathbf{I}_{1:t-1}^{c=C}) \tag{7}$$

On further factorizing the image likelihood term $p(\mathbf{I}_t^{c=C}|\{\mathbf{X}_t^{c,i}\})$, we get:

$$p(\{\mathbf{X}_t^{c,i}\}|\mathbf{I}_{1:t}^{c=C}) \propto \overbrace{p(\mathbf{Y}_t^{c=C,i}|\mathbf{X}_t^{c=C,i})}^{Appearance\ Likelihood}\ \overbrace{p(\{\mathbf{X}_t^{c\neq C,i}\}|\mathbf{X}_t^{c=C,i})}^{Spatial\ Likelihood}$$
$$\overbrace{p(\{\mathbf{X}_t^{c,i}\}|\mathbf{I}_{1:t-1}^{c=C})}^{Prior} \tag{8}$$

We propose a novel strategy to model the appearance likelihood in a distributed manner by taking viewpoint shift into account (explained in section IV-A). Also, in section IV-B we show the manner in which we combine local and global motion models using the interacting Markov Chain Monte Carlo. Most importantly, we induce scene specific priors into the global particle filter. Figure 2 shows the proposed distributed tracking methodology.

A. Appearance Modeling

We use a discriminative classifier to model the appearance of the object. The likelihood of a pixel x belonging to the foreground label $y = 1$ is given by:

$$p(y=1|x) = \frac{1}{1+e^{-\mathbf{H}_t(\mathbf{F}(x))}} \tag{9}$$

where $\mathbf{H}_t$ is a discriminative classifier learnt with online Boosting and $\mathbf{F}$ is the image feature computed at pixel location x. Given a confidence map of the object of interest by testing appearance classifiers at time t, a mean-shift based [12] object state estimate ν_t is obtained. The appearance likelihood is given as follows:

$$p(\mathbf{Y}_t|\mathbf{X}_t) = p_{pos}(\nu_t|\mathbf{X}_t) \tag{10}$$

where $p_{pos}(\nu_t|\mathbf{X}_t)$ is drawn from the Normal distribution such that $p_{pos}(\nu_t|\mathbf{X}_t) \sim \mathcal{N}(\nu_t; \mathbf{X}_t, \Sigma_{pos})$ and Σ_{pos} is the covariance matrix. In the rest of this section, we discuss in detail about the manner in which we train the local and global appearance classifiers, $\mathbf{H}_t^{(l)}$ and $\mathbf{H}_t^{(g)}$ respectively.

1) Local Appearance Modeling: For learning the local appearance classifier $\mathbf{H}_t^{(l)}$, we use a combination of histogram of oriented gradients features (HOG) [17] and normalized color features. We use a 12 dimensional feature vector with 9 bins for HOG and 3 for normalized pixel RGB color values. Similar to Ensemble Tracker [12], we train a discriminative online classifier at each time instance by using samples within the object bounding box as positive examples and samples outside the object bounding box as negative examples.

2) Global Appearance Modeling: Similar to local appearance modeling, we extract an n-dimensional feature vector with HOG and normalized pixel RGB color values (we used 36 bins for HOG and 3 for normalized pixel RGB color values). We propose a novel strategy to extract training samples for learning the global discriminative appearance model in a distributed manner. At each time instance, each tracker in every view (one tracker per object) shares their training examples to their neighboring camera views.

We first formulate the problem for two views and subsequently extend to multiple views. Let $X_1 \in \mathbb{R}^{N_1\times n}$ and $X_2 \in \mathbb{R}^{N_2\times n}$ be the set of training samples obtained from two views where N_1 and N_2 are the number of training samples in each of the views. $S_1 \in \mathbb{R}^{N_1\times d}$ and $S_2 \in \mathbb{R}^{N_2\times d}$ represent generative subspaces obtained by performing Principal Component Analysis (PCA) on X_1 and X_2 where $d << n$ (in our experiments, we set $d = 17$). We now discuss the computation of intermediate subspaces $S_m, m \in \mathbb{R}, 1 < m < 2$ in order to model the appearance changes due to viewpoint variations.

The space of d-dimensional subspaces containing the origin in $\mathbb{R}^n$ can be identified with the Grassmann manifold $\mathbb{G}_{n,d}$ where S_1 and S_2 are points on $\mathbb{G}_{n,d}$. The Grassmann manifold is the space of d dimensional subspace in $\mathbb{R}^n$ and a point on Grassmann manifold represents a subspace. We use geodesic paths that are constant velocity curves on a manifold to obtain intermediate subspaces. Viewing $\mathbb{G}_{n,d}$ as quotient space of $SO(n)$, the geodesic path starting from S_1 is given by one-parameter exponential flow $\psi(m') = Qexp(m'B)$ [1], [18] where $0 < m' < 1$ and B is restricted to a skew-symmetric, block diagonal matrix of the form:

$$B = \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix}, A \in \mathbb{R}^{(n-d)\times d}. \tag{11}$$

where A specifies the direction and the speed of geodesic flow. $Q \in SO(n)$ such that $Q^T S_1 = J$ and $J = \begin{bmatrix} I_d \\ 0_{n-d,d} \end{bmatrix}$. I_d is a $d \times d$ identity matrix. By varying m' between 0 and 1, we get a set of intermediate subspace S' that includes S_1 and S_2 (interested readers refer to [1] for more details on calculating A).

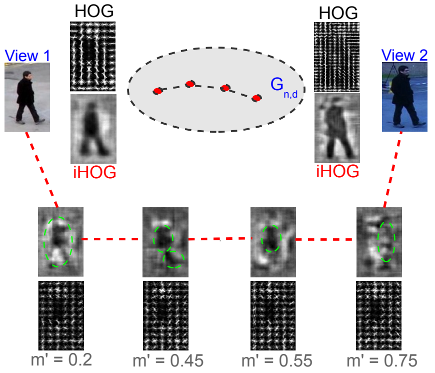

Fig. 3: **Training Sample Generation:** Training samples for global appearance learning is obtained by projecting samples from view 1 on to different generative subspaces obtained by varying m'. As highlighted in inverse HOG representation (iHOG) obtained using [19], body of the person is clearly discriminative with respect to viewpoint changes.

Given the intermediate subspaces S', we propose a novel methodology to extract training samples to learn the multi-view discriminative appearance model. Let N' be the number of intermediate subspaces. We extract training samples by projecting X_1 on to these intermediate subspaces. We train our global discriminative appearance model by gathering each of the training samples projected on to these intermediate subspaces i.e., $X_1' \in \mathbb{R}^{N'N1\times d}$. For multiple views, we adopt similar strategy to extract training samples between the current view and every other view using intermediate subspaces. Finally, we random sample training examples to learn the global appearance classifier.

B. Interacting Markov Chain Monte Carlo Based Tracking

We now present a camera tracking-by-detection algorithm with an interacting MCMC framework. For every object, we use two kinds of particle filters, local and global, with MCMC sampling. For local particle filters, the observation likelihood is computed using the local appearance model and object interaction that are local to the camera. For global particle filters, the observation likelihood is computed based on global appearance model, multi-camera information and scene priors. Similar to [20], each filter operates either in parallel or interactive mode. In the parallel mode, object state is completely determined by the local particle filter. In the interacting mode, the local particle filter interacts with the global filter and determines the object state at that time instance. By separating the local and global models, we can easily account for higher level constraints into the global model and reduce the complexity of the local particle filters significantly.

C. MCMC based Local Particle Filter

The local particle filter captures the local information within the image plane and it does not take scene priors into account for modeling the object motion. The observation model for the local particle filter is given by:

$$p_l(\mathbf{Y}_t^{c=C,i}|\mathbf{X}_t^{c=C,i}) = \overbrace{p(\mathbf{Y}_t^{c=C,i}|\mathbf{X}_t^{c=C,i})}^{Appearance\ with\ \mathbf{H}_t^{(l)}} \overbrace{\prod_{j\neq i}\psi(\mathbf{X}_t^{c=C,i},\mathbf{X}_t^{c=C,j})}^{Interaction} \tag{12}$$

$\psi(\mathbf{X}_t^{c=C,i},\mathbf{X}_t^{c=C,j})$ encodes the interaction likelihood based on the state of the other objects in the scene and it is based on the Magnetic repulsion potential:

$$\psi(\mathbf{X}_t^{c=C,i},\mathbf{X}_t^{c=C,j}) = 1 - \frac{1}{\beta}\exp(-\frac{dist(\mathbf{X}_t^{c=C,i},\mathbf{X}_t^{c=C,j})}{\sigma_i^2}) \tag{13}$$

where β is the normalization constant and σ_i^2 characterizes the allowable maximum interaction distance (in our experiments, we set $\beta = 1$ and $\sigma_i^2 = 100$). We approximate the local motion model with the Normal distribution such that $p_l(\mathbf{X}_t|\mathbf{X}_{t-1}) \sim \mathcal{N}(\mathbf{X}_t;\mathbf{X}_{t-1},\Sigma_l)$. The local particle filter finds the MAP estimate defined in equation 3 by sampling via Metropolis Hastings algorithm. The algorithm consists of two steps, a proposal step and an acceptance step. In the proposal step, the new state is proposed with the following proposal density:

$$R_l(\mathbf{X}_t^*;\mathbf{X}_t) = p_l(\mathbf{X}_t^*|\mathbf{X}_t) \tag{14}$$

where R_l denotes the proposal density function based on the local particle filter's motion model and $\mathbf{X}_t^*$ represents the new state proposed by R_l at time t. Given the proposed state, the local filter accepts the new state with the acceptance ratio given by:

$$\alpha_{parallel} = \min\left[1, \frac{p_l(\mathbf{Y}_t|\mathbf{X}_t^*)R_l(\mathbf{X}_t;\mathbf{X}_t^*)}{p_l(\mathbf{Y}_t|\mathbf{X}_t)R_l(\mathbf{X}_t^*;\mathbf{X}_t)}\right] \tag{15}$$

D. MCMC based Global Particle Filter

The global particle filter operates with the global appearance model and it takes scene priors and multi-camera information into account. The observation model, $p_g(\mathbf{Y}_t^{c=C,i}|\mathbf{X}_t^{c=C,i})$, for the global particle filter is given by:

$$p_g(\mathbf{Y}_t^{c=C,i}|\mathbf{X}_t^{c=C,i}) = \overbrace{p(\mathbf{Y}_t^{c=C,i}|\mathbf{X}_t^{c=C,i})}^{Appearance\ with\ \mathbf{H}_t^{(g)}} \overbrace{\prod_{k\neq C}\Gamma(\mathbf{X}_t^{c=C,i},\mathbf{X}_t^{c=k,i})}^{Multi-camera\ likelihood} \overbrace{\phi(\mathbf{X}_t^{c=C,i})}^{ScenePriors} \tag{16}$$

$\Gamma(\mathbf{X}_t^{c=C,i},\mathbf{X}_t^{c=k,j})$ encodes the multi-camera spatial likelihood and it is given by:

$$\Gamma(\mathbf{X}_t^{c=C,i},\mathbf{X}_t^{c=k,i}) = \frac{1}{\rho}\exp(-\frac{dist(\mathbf{X}_t^{c=C,i},\mathbf{X}_t^{c=k,i})}{\sigma_c^2}) \tag{17}$$

TABLE I: Mean Root Mean Square Pixel Errors on different datasets.

Datasets	OAB	OAB-PF	MS	MIL	MIL-PF	Struck	Proposed
Outdoor	19	20	35	19	12	13	**7**
PETS-2009	61	120	86	86	120	98	**15**
Indoor	16	11	18	17	14	22	**8**

where ρ is the normalization constant and σ_c^2 characterizes the allowable distance for multi-camera interaction (in our experiments, we set $\rho = 1$ and $\sigma_c^2 = 1000$). We share object's measurement (mean-shift estimate obtained by global appearance classifier) with the neighboring camera views for multi-camera interaction. $\phi(\mathbf{X}_t^{c=C,i})$ is based on the scene, it encodes the scene knowledge into the observation model. For modeling the scene specific priors, we used Kernel density estimator to model the probability density function on trajectories of the objects that moved over the scene for a period of time [10]. We approximate the global motion model with the Normal distribution such that $p_g(\mathbf{X}_t|\mathbf{X}_{t-1}) \sim \mathcal{N}(\mathbf{X}_t; \mathbf{X}_{t-1}, \Sigma_g)$. Similar to the local particle filter, the global particle filter uses the Metropolis Hastings algorithm based MCMC sampling.

Algorithm 1: Multi-Object Tracking with IMCMC

```
Input: $\mathbf{X}_t, \gamma$
Output: $\hat{\mathbf{X}}_t$
rand() returns a random number between 0 and 1.
if rand() $< \gamma$ then
    Accept the new state with the probability (18)
else
    Propose the new state using (14)
    Accept the new state with the probability (15)
end
Estimate the MAP state $\hat{\mathbf{X}}_t$ using (3)
```

E. IMCMC based Tracking Algorithm

During the sampling process, at each time step, the local particle filter interacts with the global particle filter (illustrated in Fig.2). We make use of the IMCMC framework for the local particle filter to communicate with the global particle filter [6], [11]. However, the global particle filter operates entirely in parallel mode. The proposed tracking algorithm operates in either parallel or interactive mode [20]. The local and global particle filters act as the parallel Metropolis Hastings in the parallel mode. Whereas in the interaction mode, the local particle filter communicates with the global particle filter and seeks a better state for the object configuration. The local particle filter then accepts the state of the global particle filter as its own state with the probability as given by:

$$\alpha_{interacting} = \frac{p_g(\mathbf{Y}_t^{c=C,i}|\mathbf{X}_t^{c=C,i})}{p_l(\mathbf{Y}_t^{c=C,i}|\mathbf{X}_t^{c=C,i}) + p_g(\mathbf{Y}_t^{c=C,i}|\mathbf{X}_t^{c=C,i})} \tag{18}$$

Algorithm 1 explains the proposed multi-camera tracking methodology with IMCMC framework. The MAP estimate is obtained using the local particle filter by communicating with the global particle filter:

TABLE II: Mean Pascal VOC detection scores on different datasets.

Datasets	OAB	OAB-PF	MS	MIL	MIL-PF	Struck	Proposed
Outdoor	0.42	0.41	0.26	0.46	0.51	0.55	**0.68**
PETS-2009	0.38	0.10	0.23	0.40	0.10	0.20	**0.66**
Indoor	0.52	0.54	0.45	0.49	0.56	0.49	**0.66**

F. Occlusion Handling

In some scenarios, objects might not be detected due to various reasons such as lighting changes, illumination effects, and missing features. We assume that an object is occluded when ζ falls below a certain threshold (0.1 in our experiments), where ζ of the image patch positioned at $\mathbf{X}_t$ is given by:

$$\zeta = \frac{sum\ of\ pixel\ likelihoods}{area\ of\ the\ patch\ positioned\ at\ \mathbf{X}_t}$$

For a given tracker, if the object is occluded, the global particle filter proceeds with the prediction step and performs an update with multi-camera information and scene priors. On the other hand, the local particle filter operates completely in interactive mode. The tracker pauses its operation after a specified number of continuously missed detections between frames (in our experiments, we set the threshold to 20 frames).

V. Experiments

We evaluated the proposed method with a wide area camera network consisting of six cameras on an outdoor environment. Videos (640x480) are captured for several hours in an uncontrolled environment with complex shape and appearance changes in human body, wireless packet losses and irregular illumination variations (for example shadows, lighting changes due to sun rays and others). Also, we evaluated the proposed algorithm in some of the publicly available multi-camera pedestrian datasets [21] and [9]. Tables I and II show average VOC detection scores and root mean square pixel errors for various algorithms on three different datasets (Outdoor, PETS2009, and Indoor). In all the experiments, we set the number of weak classifiers for the Ensemble training to 12 and updated 2 weak classifiers at every frame. We set the number of particles $P = 500$. For both local and global particle filters, we used the Brownian motion model with standard deviation $\sigma_x = 21, \sigma_y = 3$ and $\sigma_s = 0.01$. We set the IMCMC interaction threshold $\gamma = 0.1$. We assumed Σ_{pos} to be a diagonal matrix with leading diagonals equal to $[2, 2]$. Finally, for generating intermediate subspaces we set $m = [0.25, 0.5, 0.75]$. For comparison metrics, we used Root Mean Square Pixel error and pascal VOC detection score (primarily used for comparing single camera appearance based trackers). VOC detection scores capture compactness of the predicted bounding box is with respect to the ground plane bounding box. With the predicted bounding box B_p and the ground truth bounding box B_{gt}, the pascal VOC detection score is given by:

$$VOC\ \ Detection\ \ Score = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{19}$$

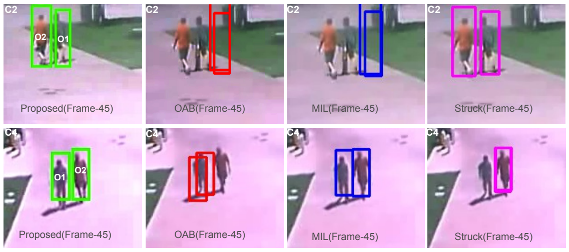

Fig. 4: **Experiment 1:** On camera C2, object O1 is indistinguishable from the background and hence some of the discriminative appearance model based trackers fail. Whereas in the proposed methodology, training samples are extracted from intermediate subspaces between multiple views to learn the global appearance model. Additionally, multi-camera information and scene priors are helpful in maneuvering the object when appearance cues are misleading. Best viewed in color.

A. *Global Appearance Modeling with Global Filtering*

In our first set of experiments with our own dataset, we evaluated the proposed methodology with six cameras in the network, of which only two of the cameras (C2 and C4) directly sensed the objects of interest (O1 and O2). The objects of interest are initialized using a background subtraction based blob detector. We compared the proposed methodology with some of the state of the art tracking algorithms: on-line Adaboost (OAB) [22], on-line Adaboost with particle filter (OAB-PF), Multiple instance learning(MIL) [13], Multiple instance learning with particle filter (MIL-PF), Struck [23], and Mean-shift (MS) [24] trackers. From the outdoor scenario results in Tables I and II, we show that the proposed methodology outperforms all the other tracking methodologies in network wide scenarios. Poor performance of some of the appearance based trackers is due to mistaken identities and inadequate appearance modeling. In the case of the proposed algorithm, a robust global appearance classifier is learnt by taking viewpoint shift into account. Additionally, the multi-camera likelihood and scene priors in global particle filters are helpful in correcting the local particle filters whenever the local particle filter suffers from insufficient motion and appearance modeling. When the object merges with background (shown in Figure 4), scene priors in global particle filters are helpful in maneuvering the object using the prior information from previous objects that moved on the same location over the time and also the multi-camera likelihood is helpful in correcting the object state estimate by exploiting multi-view redundancy. Most importantly, the global appearance model captures the missing features using the training samples obtained from neighboring camera views and helps the tracker to recover itself in the subsequent frames by building a robust appearance model.

B. *Global Appearance Modeling with Scene Priors*

For the second set of experiments with PETS-2009 sequences, the object of interest (O1) was initialized in the first (C1), third (C3) and sixth (C6) views respectively. For this experiment, to study the efficacy of the global appearance modeling with scene priors, the interactive likelihood in local particle filter was turned off. As seen in Tables I and II, the proposed algorithm outperforms state of the art algorithms on VOC detection scores and root mean square pixel errors. The success of the proposed algorithm (as illustrated in Figure 5) on this scenario could be attributed to the following: The proposed algorithm is robust to intermittent tracker failures due to the partial occlusions by learning a robust global appearance model. In this scenario, the lamp pole caused a partial occlusion in view C1 and most of the tracking algorithms failed completely due to improper appearance modeling. Similar to the first set of experiments, scene priors and multi-camera information is helpful in correcting the local particle filters during insufficient motion modeling.

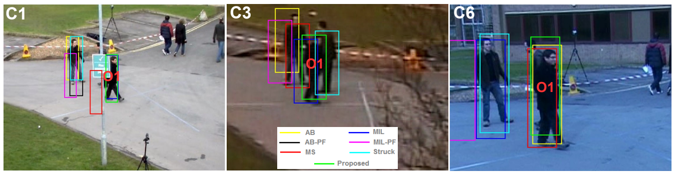

Fig. 5: **Experiment 2:** Tracking results from individual camera views are shown. Clearly, multi-camera information and scene priors help in avoiding failures due to partial occlusions. In this experiment, most of the other tracking methods failed due to the partial occlusion caused by the lamp pole. Object of interest is marked by the label 'O1'. Best viewed in color.

C. *Global Appearance Modeling without Scene Priors*

For the third set of experiments, we used indoor video sequences from [9]. The camera network consists of five cameras along a long corridor. In order to increase the difficulty of the experiment, we used a part of the sequence where the corridor lights were switched off. Also, scene priors were not available for this network. As seen in Tables I and II, the proposed algorithm outperforms the rest due to the following reasons: a) Objects appeared similar due to the lighting conditions, the global appearance model is helpful in capturing discriminative features using the training samples from neighboring views. Other methods fail to capture these variations and hence lose track of the objects due to improper appearance based association, b) Also, the interactive likelihood in the local particle filter is helpful in maintaining spatial relationship between the objects between the frames. This helps in proper association in the presence of outliers due to missing features between the frames. Missing features are more likely to happen due to irregular lighting conditions.

D. *Comparison With Multiple Camera Trackers*

We compared the proposed algorithm with state of the art multiple camera distributed filtering based tracking algorithms: Joint Probabilistic Data association with Kalman consensus filter (JPDA-KCF) [25], [26], Information consensus filtering with nearest-neighbor data association (ICF-NN) [5], Information consensus filtering with ground truth data association (ICF-GT) and Multi-camera information consensus (MTIC) [3]. In ICF-NN, the nearest observation is associated with the existing tracklet based on Hungarian algorithm. We used Struck [23] (an appearance based tracker) to generate image based observations. Ground plane estimates are obtained using homographic transformation and they serve as measurements for ground distributed fusion algorithms. Tables III show mean error (in meters) and error standard deviation for different algorithms in outdoor/indoor sequences respectively. For the indoor sequences, scene priors were not available. The proposed algorithm clearly outperforms other multi-camera tracking algorithms due to the following: a)

TABLE III: Multiple Camera Tracking Comparison in Outdoor/Indoor Sequences (Multiple Objects)

Algorithm	Mean Error (m)	Error Standard Deviation (m)
Proposed	**0.27 / 0.41**	0.23 / 0.39
MTIC	6.40 / 0.69	1.54 / 0.44
ICF-NN	25.7 / 0.56	17.4 / 0.41
JPDA-KCF	52.5 / 0.70	34.3 / 0.46
ICF-GT	12.8 / 0.53	1.3 / 0.40

Multi-view appearance model effectively captures viewpoint variations. b) Ground plane fusion in global particle filters is efficiently fed back to the local particle filters through an IMCMC approach.

VI. Conclusion

This paper presented a robust multi-camera tracking algorithm using interacting Markov Chain Monte Carlo. The tracking algorithm is formulated as a global Bayesian estimation problem and solved in a distributed manner. We proposed a novel algorithm to learn a discriminative multi-view appearance model by sharing samples across the views. We provided an efficient algorithm to combine local and global models into one using a unified probabilistic framework. The distributed multi-camera tracker takes approximately one second perframe with the Matlab implementation on a machine with 8 GB RAM and 2.67 GHZ processor. The proposed algorithm is tested on some challenging datasets and validated with objective results. As a future work, we plan to add complex crowd behavioral model into the local object interaction likelihood and account for static scene components such as entry/exit locations into the scene priors.

Acknowledgments

This work was supported by ONR grant #N00014-12-1-0503.

References

[1] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 999–1006. 1, 3

[2] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera People Tracking with a Probabilistic Occupancy Map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, 2008. 1

[3] A. T. Kamal, J. A. Farrell, and A. K. Roy-Chowdhury, "Information consensus for distributed multi-target tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, 2013. 1, 6

[4] A. T. Kamal, C. Ding, B. Song, J. A. Farrell, and A. Roy-Chowdhury, "A generalized kalman consensus filter for wide-area video networks," in *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*. IEEE, 2011, pp. 7863–7869. 1

[5] A. T. Kamal, J. A. Farrell, and A. K. Roy-Chowdhury, "Information weighted consensus," in *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*. IEEE, 2012, pp. 2732–2737. 1, 6

[6] J. Corander, M. Ekdahl, and T. Koski, "Parallell interacting mcmc for learning of topologies of graphical models," *Data mining and knowledge discovery*, vol. 17, no. 3, pp. 431–456, 2008. 1, 5

[7] C. Ding, B. Song, A. Morye, J. Farrell, and A. Roy-Chowdhury, "Collaborative sensing in a distributed ptz camera network," *Image Processing, IEEE Transactions on*, vol. 21, no. 7, pp. 3282 –3295, july 2012. 2

[8] W. Qu, D. Schonfeld, and M. Mohamed, "Distributed Bayesian Multiple-Target Tracking in Crowded Environments Using Multiple Collaborative Cameras," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 21–35, Jan. 2007. 2

[9] Z. Ni, S. Sunderrajan, A. Rahimi, and B. Manjunath, "Distributed particle filter tracking with online multiple instance learning in a camera sensor network," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, sept. 2010, pp. 37 –40. 2, 5, 6

[10] I. Saleemi, K. Shafique, and M. Shah, "Probabilistic modeling of scene dynamics for applications in visual surveillance," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 8, pp. 1472 –1485, aug. 2009. 2, 5

[11] S. Sunderrajan, S. Karthikeyan, and B. Manjunath, "Robust multiple object tracking by detection with interacting markov chain monte carlo," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, sept. 2013. 2, 5

[12] S. Avidan, "Ensemble tracking," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, june 2005, pp. 494 – 501 vol. 2. 2, 3

[13] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 983–990. 2, 6

[14] Z. Ni, S. Sunderrajan, A. Rahimi, and B. Manjunath, "Particle filter tracking with online multiple instance learning," in *International Conference on Pattern Recognition*, Aug. 2010. 2

[15] P. M. Roth, C. Leistner, A. Berger, and H. Bischof, "Multiple instance learning from multiple cameras," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 17–24. 2

[16] Z. Khan, T. Balch, and F. Dellaert, "Mcmc-based particle filtering for tracking a variable number of interacting targets," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 11, pp. 1805 –1819, nov. 2005. 2

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, june 2005, pp. 886 –893 vol. 1. 3

[18] K. A. Gallivan, A. Srivastava, X. Liu, and P. Van Dooren, "Efficient algorithms for inferences on grassmann manifolds," in *Statistical Signal Processing, 2003 IEEE Workshop on*. IEEE, 2003, pp. 315–318. 3

[19] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba, "Inverting and visualizing features for object detection," *Arxiv preprint cs.CV/1212.2278*, 2012. 4

[20] J. Kwon and K. Lee, "Visual tracking decomposition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1269–1276. 4, 5

[21] J. Ferryman and A. Shahrokni, "Pets2009: Dataset and challenge," in *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*, dec. 2009, pp. 1 –6. 5

[22] H. Grabner and H. Bischof, "On-line boosting and vision," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, june 2006, pp. 260 – 267. 6

[23] S. Hare, A. Saffari, and P. Torr, "Struck: Structured output tracking with kernels," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, nov. 2011, pp. 263 –270. 6

[24] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 5, pp. 564 – 577, may 2003. 6

[25] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," *Control Systems, IEEE*, vol. 29, no. 6, pp. 82–100, 2009. 6

[26] R. Olfati-Saber, "Kalman-consensus filter: Optimality, stability, and performance," in *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*. IEEE, 2009, pp. 7036–7042. 6