Segmentation and Tracking of Multiple Humans in Crowded Environments

Tao Zhao, Member, IEEE, Ram Nevatia, Fellow, IEEE, and Bo Wu, Student Member, IEEE,

Abstract—Segmentation and tracking of multiple humans in crowded situations is made difficult by inter-object occlusion. We propose a model based approach to interpret the image observations by multiple, partially occluded human hypotheses in a Bayesian framework. We define a joint image likelihood for multiple humans based on the appearance of the humans, the visibility of body obtained by occlusion reasoning, and foreground/background separation. The optimal solution is obtained by using an efficient sampling method, data-driven Markov chain Monte Carlo (DDMCMC), which uses image observations for proposal probabilities. Knowledge of various aspects including human shape, camera model, and image cues are integrated in one theoretically sound framework. We present experimental results and quantitative evaluation, demonstrating that the resulting approach is effective for very challenging data.

Index Terms—Multiple Human Segmentation, Multiple Human Tracking, Markov chain Monte Carlo

I. INTRODUCTION AND MOTIVATION

Segmentation and tracking of humans in video sequences is important for a number of applications, such as visual surveillance and human computer interaction. This has been a topic of considerable research in the recent past and robust methods for tracking isolated or small number of humans having only transient occlusion exist. However, tracking in a more crowded situation where several people are present and exhibit persistent occlusion, remains challenging. The goal of this work is to develop a method to detect and track humans in the presence of persistent and temporarily heavy occlusion. We do not require that humans be isolated, i.e. un-occluded, when they first enter the scene. However, in order to “see” a person, we require that at least the head-shoulder region must be visible. We assume a stationary camera so that motion can be detected by comparison with a background model. We do not require the foreground detection to be perfect, e.g. the foreground blobs may be fragmented, but we assume that there are no significant false alarms due to shadows, reflections, or other reasons. We also assume that the camera model is known and that people walk on a known ground plane.

Fig.1(a) shows a sample frame of a crowded environment and Fig.1(b) shows the motion blobs detected by comparison with the learned background. It is apparent that segmenting humans from such blobs is not straightforward. One blob may include multiple objects; while one object may split into multiple blobs. Blob tracking over extended periods, e.g. [20], may resolve some of these ambiguities but such approaches are likely to fail when occlusion is persistent. Some approaches have been developed to handle occlusion, e.g. [9], but require the objects to be initialized before occlusion happens. This is usually infeasible for crowded scenes. We believe that use of a shape model is necessary to achieve individual human segmentation and tracking in crowded scenes.

In earlier related work [54], Zhao and Nevatia model human body as a 3D ellipsoid and human hypotheses are proposed based on head top detection from foreground boundary peaks. This method works reasonably well in presence of partial occlusions if the number of people in the field of view is small. As the complexity of the scene grows, head tops cannot be obtained by simple foreground boundary analysis and more complex shape models are needed to fit more accurately with the observed shapes. Also, joint reasoning about the collection of objects is needed, rather than the simpler one-by-one verification method in [54]. The consequence of this joint consideration is that the optimal solution has to be computed in the joint parameter space of all the objects. To track the objects in multiple frames, temporal coherence is another desired property besides accuracy of the spatial segmentation. We adapt a data-driven Markov chain Monte Carlo approach to explore this complex solution space. To improve the computational efficiency, we use direct image features from bottom-up image analysis as importance proposal probabilities to guide the moves of the Markov chain. The main features of this work include

1) a 3-dimensional part based human body model, which enables segmentation and tracking of humans in 3D and inference of inter-object occlusion naturally;
2) a Bayesian framework which integrates segmentation and tracking based on a joint likelihood for the appearance of multiple objects;

Fig. 1. An sample frame, the corresponding motion blobs and our segmentation and tracking result for crowded situation.
3) design of an efficient Markov chain dynamics, directed by proposal probabilities based on image cues; and
4) the incorporation of a color based background model in a mean shift tracking step.

Our method is able to successfully detect and track humans in scenes of complexity shown in Fig.1 with high detection and low false alarm rates; the tracking results for the frame in Fig.1(a) is shown in Fig.1(c) (the result includes integration of multiple frames during tracking). In the result section, we give graphical and quantitative results on a number of sequences. Parts of our system have been partially described in [53] and [55]; this paper provides a unified presentation of the methodology, additional results and discussions. This approach has been built on by other researchers, e.g. [41]. The same framework has also been successfully applied to vehicle segmentation and tracking in challenging cases [43].

The rest of the paper is organized as follows: Section II gives a brief review of the related works; Section III presents an overview of our method; Section IV describes the probabilistic modeling of the problem; Section V describes our MCMC based solution; Section VI shows experimental results and evaluation; conclusions and discussions are given in the last section.

II. RELATED WORK

We summarize related work in this section; some of these are referred to in more detail in the following sections. Due to the size of the literature in this field, it is not possible for us to provide a comprehensive survey but we attempt to include the major trends.

The observations for human hypotheses may come from multiple cues. Many previous approaches [20], [9], [54], [37], [44], [15], [18], [40], [24], [3], [45] use motion blobs detected by comparing pixel colors in a frame to learned models of the stationary background. When the scene is not highly crowded, most part of the humans in the scene are detected in the foreground motion blob; multiple humans may be merged into a single blob but they can be separated by rather simple processing. For example, Haritaoglu et al. [15] uses vertical projection of the blob to help segment a big blob into multiple humans. Siebel and Maybank [40], Zhao and Nevatia [54] detect head candidates by analyzing the foreground boundaries. Since different humans have small overlapping foreground regions, they could be segmented in a greedy way. However, the utility of these methods in crowded environments such as in Fig.1 is likely to be limited.

Some methods, e.g. [50], [31], [7], [13] detect appearance- or shape-based patterns of humans directly. [50] and [31] learn human detectors from local shape features; [7] and [13] builds contour templates for pedestrians. These learning based methods need a large number of training samples and may be sensitive to imaging view-point variations as they learn 2-D patterns. Besides motion and shape, face and skin-color are also useful cues for human detection, but environments where these cues could be utilized are limited, usually indoor scenes where illumination is controlled and the objects are imaged with high resolution, e.g. [42], [12].

Without a specific model of objects, tracking methods are limited to blob tracking e.g. [3]. The main advantage of model-based tracking is that it can solve the blob merge and split problems by enforcing a global shape constraint. The shape models could be either parametric, e.g. an ellipsoid as in [54], or non-parametric, e.g. the edge template as in [13]: either in 2D, e.g. [46] or in 3D, e.g. [54]. Parametric models are usually generative and of high dimensionality, while non-parametric models are usually learned from real samples. 2D models make the matching of hypotheses and image observations straightforward, while 3D models are more natural for occlusion reasoning. The choice of the model complexity depends on both the application and the video resolution. For human tracking from a mid-distant camera, we do not need to capture the detailed body articulation, a rough body model, such as the generic cylinder in [19], the ellipsoid in [54], and the multiple rectangles in [46] suffice. When the body pose of humans is desired and the video resolution is high enough, more complex models could be used, such as the articulated models in [54] and [34].

Tracking of multiple objects requires matching of hypotheses with the observations both spatially and temporally. When objects are highly inter-occluded, their image observations are far from being independent, hence a joint likelihood for multiple objects is necessary [46], [27], [19], [35], [30], [51]. Smith et al. [41] use a pair-wise Markov Random Field (MRF) to model the interaction between humans and define the joint likelihood. Rittscher et al. [36] include a hidden variable, which indicates a global mapping from the observed features to human hypotheses, in the state vector.

As the solution space is of high dimension, searching for the best interpretation by brute force is not feasible. Particle filters based methods, e.g. [19], [46], [30], [51], [27], become unsuitable when the dimensionality of the search space is high as the number of samples needed usually grows exponentially with the dimension. [41], [21] use some variations of MCMC algorithm to sample the solution space while [45], [36] uses an EM style method. For efficiency the candidate solutions could be generated from some image cues, not pure randomly, e.g. [36] propose hypotheses from local silhouette features.

Information from multiple cameras with overlapping views can reduce the ambiguity of a single camera. Such methods usually assume that at least from one view point, the object can be detected successfully (e.g. [11]) or many cameras are available for 3-dimensional reconstruction (e.g. [28]). The difficulty of segmenting multiple humans which overlap in images from a stereo camera is alleviated by analyzing in the 3-dimensional space where they are separable [52]. In a multi-camera context, an object can be tracked even when it is fully occluded from some of the views; however, many real environments do not permit use of multiple cameras with overlapping views. In this paper, we consider situations where video from only one camera is available. However our approach can utilize multiple cameras with little modification.

MCMC-based methods are receiving increasing popularity for computer vision problems due to its flexibility in optimizing an arbitrary energy function as opposed to energy functions of specific type as in graph cut [2] or belief propagation [49].
It has been used for various applications including segmenting multiple cells [38], image parsing [48], multi-object tracking [21], estimating articulated structures [23], etc. Data-drive MCMC was proposed by [48] to utilize bottom-up image cues to speed up the sampling process.

We want to point out the difference between our approach and another independently developed work [21] which also used MCMC for multi-object tracking. [21] assumes that the objects do not overlap by applying a penalty term for overlap while our approach explicitly uses a likelihood of appearance under occlusion. Our approach focuses on the domain of tracking human which is the most important subject for visual surveillance. We consider the 3-dimensional perspective effect in typical camera setting while the ant tracking problem described in [21] is almost a 2-dimensional problem. We utilize acquired appearance where each object is of different appearance where ants in [21] are assumed to have the same appearance. We developed a full set of effective bottom-up cues for human segmentation and hypotheses generation.

III. OVERVIEW

Our approach to segmenting and tracking of multiple humans emphasizes the use of shape models. An overview diagram is given in Fig.2. Based on a background model, the foreground blobs are extracted as the basic observation. By using the camera model and the assumption that objects move on a known ground plane, multiple 3D human hypotheses are projected onto the image plane and matched with the foreground blobs. Since the hypotheses are in 3D, occlusion reasoning is straightforward. In one frame, we segment the foreground blobs into multiple humans and associate the segmented humans with the existing trajectories. Then the tracks are used to propose human hypotheses in the next frame. The segmentation and tracking are integrated in a unified framework and inter-operate along time.

![Fig. 2. Overview diagram of our approach.](image)

We formulate the problem of segmentation and tracking as one of Bayesian inference to find the best interpretation given the image observations, the prior models, and the estimates from previous frame analysis (i.e. the maximum a posteriori, MAP, estimation). The state to be estimated at each frame includes the number of objects, their correspondences to the objects in the previous frame (if any), their parameters (e.g. positions), and the uncertainty of the parameters. We define a color-based joint likelihood model which considers all the objects and the background together, and encodes both the constraints that the object should be different from the background and that the object should be similar to its correspondence. Using this likelihood model gracefully integrates segmentation and tracking, and avoids a separate, sometimes ad hoc, initialization step. Given multiple human hypotheses, before calculating the joint image likelihood inter-object occlusion reasoning is done. The occluded parts of a human should not have corresponding image observations.

The solution space contains subspaces of varying dimensions, each corresponding to a different number of objects. The state vector consists of both discrete and continuous variables. This disqualifies many optimization techniques. Therefore we use a highly general reversible jump/diffusion MCMC-based method to compute the MAP estimate. We design dynamics for multi-object tracking problem. We also use various direct image features to make the Markov chain more efficient. Direct image features alone do not guarantee optimality because they are usually computed locally or using partial cues. Using them as proposal probabilities of the Markov chain results in an integrated top-down/bottom-up approach which has both the computational efficiency of image features and the optimality of a Bayesian formulation. A mean shift technique [5] is used as efficient diffusion for the Markov chain. The data-driven dynamics and the in-depth exploration of the solution space make the approach less sensitive to dimensionality compared to particle filters. Our experiments show that the described approach works robustly in very challenging situations with affordable computation; some results are shown in Section VI.

IV. PROBABILISTIC MODELING

Let $\theta$ represent the state of the objects in the scene at time $t$; it consists of the number of objects in the scene, their 3D positions and other parameters describing their size, shape and pose. Our goal is to estimate the state at time $t$, $\theta(t)$, given the image observations, $I(1), \ldots, I(t)$, abbreviated as $I(1, \ldots, t)$. We formulate the tracking problem as computing the maximum a posteriori (MAP) estimation, $\theta(t)^*$. 

$$
\theta(t)^* = \arg \max_{\theta(t) \in \Theta} P(\theta(t) | I(1, \ldots, t)) \\
= \arg \max_{\theta(t) \in \Theta} \left\{ P(I(t) | \theta(t)) P(\theta(t) | I(1, \ldots, t-1)) \right\} 
$$

(1)

where $\Theta$ is the solution space. Denote by $m$ the state vector of one individual object. A state containing $n$ objects can be written as $\theta = \{(k_1, m_1), \ldots, (k_n, m_n)\} \in \Theta_n$, where $k_i$ is the unique identity of the $i$-th object whose parameters are $m_i$, and $\Theta_n$ is the solution space of exactly $n$ objects. The entire solution space is $\Theta = \bigcup_{n=0}^{N_{max}} \Theta_n$, where $N_{max}$ is a upper bound of the number of objects. In practice, we compute an approximation of $P(\theta(t) | I(1, \ldots, t-1))$ (details are given later in section IV-D).

A. 3D Human Shape Model

The parameters of an individual human, $m$, are defined based on a 3D human shape model. Human body is highly articulated, however, in our case, the human motion is mostly limited to standing or walking, and we do not attempt to
capture the detailed shape and articulation parameters of the human body. Thus we use a number of low dimensional models to capture the gross shape of human bodies.

![Image of 3D human models](image)

Fig. 3. A number of 3D human models to capture the gross shape of human bodies.

Ellipsoids fit human body parts well and has the property that its projection is an ellipse with a convenient form [16]. Therefore we model human shape by a composition of multiple ellipsoids corresponding to the head, the torso and the legs, with fixed spatial relationship. A few such models at characteristic poses are sufficient to capture the gross shape variations of most humans in the scene for mid-resolution images. We use the multi-ellipsoid model to control the model complexity while maintaining a reasonable level of fidelity. We have used three such models (1 for legs close to each other and 2 for legs well-split) in our previous work on multi-human segmentation [53]. However, in this work we use only a single model with three ellipsoids which we found sufficient for tracking.

The model is controlled by two parameters called size and thickness. The size parameter is the 3D height of the model; it also controls the overall scaling of the object in the three directions. The thickness parameter captures extra scaling in the horizontal directions. Besides size and thickness, the parameters also include image position of the head 1, 3D orientation of the body, and 2D inclination of the body. The orientations of the models are quantized into a few levels for computation efficiency. The origin of the rotation is chosen so that 0° corresponds to human facing the camera. We use 0° and 90° to represent frontal/back and side view in this work. The 3D models assumes that humans are perfectly upright, but there are chances that they incline their body slightly. We use one parameter to capture the inclination in 2D (as opposed to two parameters in 3D). Therefore, the parameters of the i-th human are \( \mathbf{m}_i = \{ x_i, y_i, h_i, f_i, i_i \} \) which are \( \text{orientation}, \text{position}, \text{size}, \text{thickness}, \text{and inclination} \) respectively. We also write \( (x_i, y_i) \) as \( \mathbf{u}_i \).

With a given camera model and a known ground plane, the 3D shape models automatically incorporates the perspective effect of camera projection (change in object image size and shape due to the change in object position and/or camera viewpoint). Compared to 2D shape models (e.g. [131]) or pre-learnt 2D appearance models (e.g. [50]), the 3D models are more easily applicable for a novel viewpoint.

\(^1\)The image head location is a equivalent parameterization of the world location on the ground plane \((x_w, y_w)\) given the human height. The two are related by \([x, y, 1]^T \sim [p_x, p_y, p_z + \mathbf{h}][x_w, y_w, 1]^T\), where \(p_r\) is the i-th column of the camera projection matrix and \(h\) is the height of the human. For clarity of presentation, we chose the ground plane to be \(z = 0\).

### B. Object Appearance Model

Besides the shape model, we also use a color histogram of the object, \( \hat{p} = \{ p_1, \ldots, p_m \} \) (\(m\) is the number of bins of the color histogram) defined within the object shape, as a representation of its appearance which helps establish correspondence in tracking. We use color histogram because it is insensitive to the non-rigidity of human motion. Furthermore, there exists efficient algorithm, e.g. the mean shift technique [5], to optimize a histogram-based object function. When calculating the color histogram, a kernel function \(K_E\) with Epanechnikov profile [5] is applied to weight pixel locations so that the center has a higher weight than the boundary. Such a representation has been used in [6]. Our implementation uses a single RGB histogram with 512 bins (8 for each dimension), of all the samples within the three elliptic regions of our object model.

### C. Background Appearance Model

The background appearance model is a modified version of a Gaussian distribution. Denote by \((\bar{r}_j, \bar{g}_j, \bar{b}_j)\) and \(\Sigma_j = diag\{\sigma_{r_j}^2, \sigma_{g_j}^2, \sigma_{b_j}^2\}\) the mean and the covariance of the color at pixel \(j\). The probability of pixel \(j\) being from the background is

\[
P_b(I_j) = P_b(r_j; g_j, b_j) = \max_{\epsilon} \left\{ \exp \left[- \left( \frac{r_j - \bar{r}_j}{\sigma_{r_j}} \right)^2 - \left( \frac{g_j - \bar{g}_j}{\sigma_{g_j}} \right)^2 - \left( \frac{b_j - \bar{b}_j}{\sigma_{b_j}} \right)^2 \right], \epsilon \right\}
\]

where \(\epsilon\) is a small constant. It is a composition of a Gaussian distribution and a uniform distribution. The uniform distribution captures the outliers which are not modeled by the Gaussian distribution to make the model more robust. The Gaussian parameters (mean and covariance) are updated continuously by the video stream only with the non-moving regions. More sophisticated background model (e.g. mixture of Gaussian [44] or non-parametric [10]), could be used to account for more variations but this is not the focus of this work; we assume that comparison with background model yields adequate foreground blobs.

### D. The Prior Distribution

The prior distribution \(P(\theta^{(t)} | I^{(1, \ldots, t-1)})\) is decomposed in two parts given by:

\[
P(\theta^{(t)} | I^{(1, \ldots, t-1)}) \propto P(\theta^{(t)}) P(I^{(1, \ldots, t-1)})
\]

\(P(\theta^{(t)})\) is independent of time, and is defined by

\[
\prod_{i=1}^{n} P(S_i I_{|S_i|}) P(m_i), \quad \text{where } S_i \text{ is the projected image of the } i\text{-th object and } |S_i| \text{ is its area.}
\]

The prior of the image area \(P(|S_i|)\) is modeled as being proportional to \(\exp(-\lambda_1|S_i|) [1 - \exp(-\lambda_2|S_i|)]^2\). The first term here penalizes large total object size to avoid situations where two hypotheses overlap a large portion of an image blob,

\(2\)We have used prior on the number of objects in [53] to constrain over segmentation. However we found that the prior on the area is more effective due to the large variation of the image sizes of the objects (due to camera perspective effect) and therefore their different contribution to the likelihood.
while the second term penalizes objects with small image sizes as they are more likely to be due to image noise. Although the prior on 2D image size could be converted to the 3D space, defining this prior in 2D is more natural, because these properties model the reliability of image evidence independent of the camera models. The priors on the human body parameters are considered independently. Thus we have \( P(m_i) = P(o_j)P(x_i, y_i)P(h_i)P(f_i)P(i_i). \) We set \( P(o_{forward}) = P(o_{projection}) = 1/2. \) \( P(x_i, y_i) \) is a uniform distribution in the image region where a human head is plausible. \( P(h_i) \) is a Gaussian distribution \( N(\mu_h, \sigma_h^2) \) truncated in the range of \([h_{min}, h_{max}]\) and \( P(f_i) \) is Gaussian distribution \( N(\mu_f, \sigma_f^2) \) truncated in the range of \([f_{min}, f_{max}]\). \( P(i_i) \) is a Gaussian distribution \( N(\mu_i, \sigma_i^2) \). In our experiments, we use \( \mu_h = 1.7m, \sigma_h = 0.2m, h_{min} = 1.5m, h_{max} = 1.9m; \) \( \mu_f = 1, \sigma_f = 0.2, f_{min} = 0.8, f_{max} = 1.2; \) \( \mu_i = 0, \sigma_i = 3^\circ. \) These parameters correspond to common adult body sizes.

We approximate the second term of the right side of Equ.3, \( P(\theta(t)|I^{t-1}) \), by \( P(\theta(t)|\theta(t-1)) \), assuming \( \theta(t) \) encodes the necessary information from the past observations. For convenience of expression, we rearrange \( \theta(t) \) and \( \theta(t-1) \) as \( \tilde{\theta}(t) = \left\{ \tilde{k}_i(t), \tilde{m}_i(t) \right\}_{i=1}^N \) and \( \tilde{\theta}(t-1) = \left\{ \tilde{k}_i(t-1), \tilde{m}_i(t-1) \right\}_{i=1}^N \), where \( N \) is the overall number of object present in the two frames, so that one of \( \tilde{k}_i(t) = 1 \) means object \( k_i(t) \) is a tracked object; \( \tilde{m}_i(t) = \phi \) means object \( m_i(t) \) is a dead object (i.e. trajectory is terminated); and \( \tilde{m}_i(t) = \phi \) means object \( \tilde{k}_i(t) \) is a new object. With the rearranged state vector, we have \( P(\tilde{\theta}(t)|\tilde{\theta}(t-1)) = \prod_{i=1}^N P\left(\tilde{m}_i(t)|\tilde{m}_i(t-1)\right) \). The temporal prior of each object follows the definition

\[
P\left(\tilde{m}_i(t)|\tilde{m}_i(t-1)\right) \propto \begin{cases} P_{assoc}(\tilde{m}_i(t)|\tilde{m}_i(t-1)), & \tilde{k}_i(t) = \tilde{k}_i(t-1) \\ P_{new}(\tilde{m}_i(t)), & \tilde{k}_i(t) = \tilde{k}_i(t-1) = 1 \\ P_{dead}(\tilde{m}_i(t-1)), & \tilde{m}_i(t) = \phi \\ \end{cases}
\]

We assume that the position and the inclination of an object follow constant velocity models with Gaussian noise, and that the height and thickness follow a Gaussian distribution (for simplicity of presentation, we omit the velocity terms in the state). We use Kalman filters for temporal estimation; \( P_{assoc} \) is therefore a Gaussian distribution. \( P_{new}(\tilde{m}_i(t)) = P_{new}(\tilde{u}_i(t)) \) and \( P_{dead}(\tilde{m}_i(t-1)) = P_{dead}(\tilde{u}_i(t-1)) \) are the likelihoods of the initialization of a new track at position \( u_i(t) \) and the termination of an existing track at position \( u_i(t-1) \) respectively. They are set empirically according to the distance of the object to the entrances/exits (the boundaries of the image and other areas that people move in/out). \( P_{new}(u) \sim N(\mu(u), \Sigma_u) \), where \( \mu(u) \) is the location of the closest entrance point to \( u \) and \( \Sigma_u \) is its associated covariance matrix which is set manually or through a learning phase. \( P_{dead}(\cdot) \) follows a similar definition.

### E. Joint Image Likelihood for Multiple Objects and Background

The image likelihood \( P(I|\theta) \) reflects the probability that we observe image \( I \) (or some features extracted from \( I \)) given state \( \theta \). Here we develop a likelihood model based on the color information of background and objects. Given a state vector \( \theta \), we partition the image into different regions corresponding to different objects and the background. Denote by \( S_i \) the visible part of the \( i \)-th object defined by \( m_i \). The visible part of an object is determined by the depth order of all the objects, which can be inferred from their 3D positions and the camera model. The entire object region \( S = \cup_{i=1}^{n} S_i = \sum_{i=1}^{n} S_i \), since \( S_i \) are disjoint regions. We use \( S \) to denote the supplementary region of \( S \), i.e. the non-object region. The relationship of the regions is illustrated in Fig.4.

![Fig. 4. First pane: the relationship of visible object regions and the non-object region. Rest panes: the color likelihood model. In \( S \), the color likelihood favors both the difference of an object hypothesis with the background and its similarity with its corresponding object in a previous frame. In \( S \), the likelihood penalizes the difference with the background model. Note that the elliptic models are used for illustration.](image)

In case of multiple objects which can possibly overlap in the image, the likelihood of the image given the state cannot be simply decomposed into the likelihood of each individual objects. Instead, a joint likelihood of the whole image given all objects and the background model needs to be considered. The joint likelihood \( P(I|\theta) \) consists of two terms corresponding to the object region and the non-object region

\[
P(I|\theta) = P(I^S|\theta) P(I^\bar{S}|\theta)
\]

After obtaining \( \hat{S}_i \) by occlusion reasoning, the object region likelihood can be calculated by

\[
P(I^S|\theta) = \prod_{i=1}^{n} P(I^S|\tilde{m}_i)
\]

\[
\propto \exp \left\{ \lambda_S \sum_{i=1}^{n} |\hat{S}_i| \left[ -\lambda_B B(p_i, \tilde{p}_i) + \lambda_B B(p_i, \tilde{p}_j) \right] \right\}
\]

where \( \lambda_S \) is the color histogram of the background region within the visibility mask of object \( i \), \( \tilde{p}_i \) is the color histogram of the object, both weighted by the kernel function \( K_E(\cdot) \). \( B(p, d) = \sum_{j=1}^{m} p_j d_j \) is the Bhattachayya coefficient, which reflects the similarity of two histograms.

This likelihood favors both the difference of an object hypothesis with the background and its similarity with its corresponding object in a previous frame (Fig.4). This enables simultaneous segmentation and tracking in the same object function. We call the two terms background exclusion and
object attraction respectively. The background exclusion concept was also proposed by [33]. $\lambda_b$ and $\lambda_f$ weight the relative contribution of the two terms (we constrain $\lambda_b + \lambda_f = 1$). The object attraction term is the same as the likelihood function used in [6]. For an object without a correspondence, i.e. a new object, only the background exclusion part is used.

The non-object probability is calculated by

$$
P(I^S|\theta) = \prod_{j \in S} (P_b(I_j))^\lambda_b \propto \exp \left(-\sum_{j \in S} e_j\right),
$$

(7)

where $e_j = \log(P_b(I_j))$ is the probability of belonging to the background model, as defined in Equation 2. $\lambda_b$ in Equation 6 and $\lambda_S$ in Equation 7 weight the balance of the foreground and the background considering the different probabilistic models being used. The posterior probability is obtained by combining the prior, Equation 3, and the likelihood, Equation 5.

V. COMPUTING MAP BY EFFICIENT MCMC

Computing the MAP is an optimization problem. Due to the joint consideration of an unknown number of objects, the solution space contains subspace of varying dimensions. It also includes both discrete variable and continuous variables. These has made the optimization challenging. We use a Markov chain Monte Carlo method with jump/diffusion dynamics to sample the posterior probability. Jumps cause the Markov chain to move between subspaces with different dimensions and traverse the discrete variables; diffusions make the Markov chain sample continuous variables. In the process of sampling, the best solution is recorded and the uncertainty associated with the solution is also obtained.

Fig.5 gives a block diagram of the computation process. The MCMC based algorithm is an iterative process, starting from an initial state. In each iteration, a candidate is proposed from the state in the previous iteration assisted by image features. The candidate is accepted probabilistically according to the Metropolis-Hasting rule [17]. The state corresponding to the maximum posterior value is recorded and becomes the solution.

Suppose we want to design a Markov chain with stationary distribution $P(\theta) = P(\theta^{(t)}|I^{(t)}, \theta^{(t-1)})$. At the $g$-th iteration, we sample a candidate state $\theta'$ according to $\theta_{g-1}$ from a proposal distribution $q(\theta|\theta_{g-1})$. The candidate state $\theta'$ is accepted with the probability $p = \min\left\{1, \frac{P(\theta') q(\theta_{g-1}|\theta')}{P(\theta_{g-1}) q(\theta'|\theta_{g-1})}\right\}$.

If the candidate state $\theta'$ is accepted, $\theta_g = \theta'$, otherwise, $\theta_g = \theta_{g-1}$. It can be proven that the Markov chain constructed in this way has its stationary distribution equal to $P(\cdot)$, independent of the choice of the proposal probability $q(\cdot)$ and the initial state $\theta_0$ [47]. However, the choice of the proposal probability $q(\cdot)$ can affect the efficiency of the MCMC significantly. Random proposal probabilities will lead to very slow mixing rate. Using more informed proposal probabilities, e.g. as in data-driven MCMC [48], will make the Markov chain traverse the solution space more efficiently. Therefore the proposal distribution is written as $q(\theta, |\theta_{g-1}, I)$. If the proposal probability is informative enough so that each sample can be thought of as a hypothesis, then the MCMC approach becomes a stochastic version of the hypothesis and test approach.

In general, the original version of MCMC has dimension matching problem for solution space with varying dimensionality. A variation of MCMC, called trans-dimensional MCMC [14] is proposed to solve this problem. However, with some appropriate assumption and simplification, trans-dimensional MCMC can be reduced to the standard MCMC. We address this issue later in this section.

A. Markov Chain Dynamics

We design the following reversible dynamics for the Markov chain to traverse the solution space. The dynamics corresponding to the proposal distribution with a mixture density $q(\theta|\theta_{g-1}, I) = \sum_{a \in A} p_a q_a(\theta|\theta_{g-1}, I)$, where $A$ is the set of all dynamics $= \{\text{add}, \text{remove}, \text{establish}, \text{break, exchange, diff}\}$. The mixing probabilities $p_a$ are the chances of selecting different dynamics and $\sum_{a \in A} p_a = 1$.

We assume that we have the sample in the $g-1$th iteration $\theta_{g-1} = \{(k_1, m_1), \ldots, (k_n, m_n)\}$ and now propose a candidate $\theta'$ for the $g$-th iteration ($t$ is omitted where there is no ambiguity).

**Object hypothesis addition** Sample the parameters of a new human hypothesis $(k_{n+1}, m_{n+1})$ and add it to $\theta_{g-1}$. $q_{\text{add}}(\theta_{g-1} \cup \{(k_n, m_n)\}|\theta_{g-1}, I) = 1/n$. If $k_n$ has a correspondence in $\theta_{g-1}$, then that object becomes dead.

**Establish correspondence** Randomly select a new object $r$ in $\theta_{g-1}$ and a dead object $r'$ in $\theta_{g-1}$, and establish their temporal correspondence. $q_{\text{establish}}(\theta'|\theta_{g-1}) \propto \|u_r - u_{r'}\|^2$ for all the qualified pairs.

**Break correspondence** Randomly select an object $r$ where...

---

\[\text{Base on our experiments, we find that approximating the ratio in the second term with just the posterior probability ratio, } \frac{P(\theta') q(\theta_{g-1}|\theta')}{P(\theta_{g-1}) q(\theta'|\theta_{g-1})}, \text{ gives almost the same results as the complete computation, hence we use this approximation in our implementation.}\]
with a uniform distribution and change \( k_r \) to a new object (and same object in \( \theta^{(t-1)} \) becomes dead). 

\( q_{\text{break}} (\theta^{i} | \theta_{g-1}) = 1/n' \), where \( n' \) is the number of objects in \( \theta_{g-1}^{(t-1)} \) that have correspondences in the previous frame.

**Exchange identity** Exchange the IDs of two close-by objects. Randomly select two objects \( r_1, r_2 \in [1, n] \) and exchange their IDs. 

\( q_{\text{exchange}} (r_1, r_2) \propto ||u_{m_r_1} - u_{m_r_2}||^{-2} \). Identities exchange can also be replaced by the composition of breaking and establishing correspondence. It is used to ease the traversal since breaking and establishing correspondences may lead to a big decrease in the probability and are less likely to be accepted.

**Parameter update** Update the continuous parameters of an object. Randomly select an existing human hypothesis \( r \in [1, n] \) with a uniform distribution, and update its continuous parameters \( q_{d_{eff}} (\theta^{i} | \theta_{g-1}) = (1/n) q_d (m^{r}_{i}, m_{r}) \).

Among the above, addition and removal are a pair of reverse moves, as are the establishing and breaking correspondences; exchanging identity, and parameter updating are the their own reverse moves.

**B. Informed Proposal Probability**

In theory, the proposal probability \( q() \) does not affect the stationary distribution. However, different \( q() \) lead to different performance. The number of samples needed to get a good solution strongly depends on the proposal probabilities. In this application, the proposal probability of adding a new object, and the update of the object parameters, are the two most important ones. We use the following informed proposal probabilities to make the Markov chain more intelligent and thus have a higher acceptance rate.

**Object addition** We add human hypotheses from three cues, foreground boundaries, intensity edges, and foreground residue (foreground with the existing objects carved out). In [54] a method to detect the heads which are on the boundary of the foreground is described. The basic idea is to find the local vertical peaks of the boundary. The peaks are further verified by checking if there are enough foreground pixels below it according to a human height range and the camera model. This detector has a high detection rate and is also effective when the human is small and image edges are not reliable; however, it cannot detect the heads in the interior of the foreground blobs. Fig.6(a) shows an example of head detection from foreground boundaries.

The second head detection method is based on an “Ω” shape head-shoulder model (this term was first introduced in [53]). This detector matches the Ω shape edge template with the intensity edges to find the head candidates. First, Canny edge detector is applied to the foreground region of the input image. A distance transformation [1] is computed on the edge map. Fig.6(b) shows the exponential edge map where \( E(x, y) = \exp(-\lambda D(x, y)) \) \( D(x, y) \) is the distance to the closest edge point and \( \lambda \) is a factor to control the response field depending on the object scale in the image; we use \( \lambda = 0.25 \). Besides, the coordinates of the closest pixel point are also recorded as \( \vec{C}(x, y) \). The unit image gradient vector \( \vec{O}(x, y) \) is only computed at edge pixels. The “Ω” shape model, see Fig.6(c), is derived by projecting a generic 3D human model to the image and taking the contour of the whole head and the upper quarter torso as the shoulder. The normals of the contour points are also computed. The size of the human model is determined by the camera calibration assuming an average human height.

Denote \( \{\vec{u}_1, ..., \vec{u}_k\} \) and \( \{\vec{v}_1, ..., \vec{v}_k\} \) as the positions and the unit normals of the model points respectively when head top is at \((x, y)\). The model is matched with the image as \( S(x, y) = (1/k) \sum_{i=1}^{k} e^{-\lambda D(\vec{u}_i) (\vec{v}_i \cdot \vec{O}(\vec{C}(\vec{u}_i)))} \). A head candidate map is constructed by evaluating \( S(x, y) \) on every pixel in the dilated foreground region. After smoothing it, we find all the peaks above a threshold such that a very high detection rate but may also result in a high false alarm rate. An example is shown in Fig.6(d). The false alarms tend to happen in the area of rich texture where there are abundant edges of various orientations.

Finally, after some human objects obtained from the first two methods are hypothesized and removed from the foreground, the foreground residue map \( R = F * \mathcal{S} \) is computed. Morphological “open” operation with a vertically elongated structural element is applied to remove thin bridges and small/thin residues. From each connected component \( c \), human candidates can be generated assuming 1) the centroid of \( c \) is aligned with the center of human body; 2) the top center point of \( c \) is aligned with the human head; or 3) the bottom center point of \( c \) is aligned with the human feet.

The proposal probability for addition combines these three head detection methods \( q_{\text{a1}}(k, m) = \sum_{i=1}^{3} \lambda_{ai} q_{ai}(k, m) \), where \( \lambda_{ai} = 1, 2, 3 \) are mixing probabilities of the three methods and we use \( \lambda_{ai} = 1/3 \). \( q_{\text{ai}}() \) samples \( m \) first and then \( k \). \( q_{\text{ai}}(k, m) = q_{ai}(m)q_{ai}(k|m) \), and \( q_{\text{ai}}(m) = q_{a1}(o)q_{ai}(u)q_{ai}(h)q_{ai}(f)q_{ai}(i) \). \( q_{\text{a1}}(u) \) answers the question “where to add a new human hypothesis”. In practice, \( q_{a1}(o) \), \( q_{a1}(h) \), \( q_{a1}(f) \), and \( q_{a1}(i) \) use their respective prior distributions, and \( q_{a1}(u) \) is a mixture of Gaussian based on the bottom-up detection results. For example, denote by \( HC_1 = \{(x_i, y_i)\}_{i=1}^{N} \) the head candidates obtained by the first method, then \( q_{a1}(u) = q_{a1}(x, y) \sim \sum_{i=1}^{N} \mathcal{N}(x_i, y_i, \text{diag}(\sigma^2_x, \sigma^2_y)) \).

The definition of \( q_{a2}(u) \) and \( q_{a3}(u) \) are similar. After
u' is sampled, \( q(k|m) \propto q(k|u') \) is to sample \( k \) from 
\[ \left\{ u_{d_i}^{(t-1)}, \ldots, u_{d_i}^{(t-1),new} \right\} \]
according to \( P(u_{d_i}^{(t-1)}) \), see Equation 4, \( i = 1, \ldots, n_d \) and \( P_{new}(u) \), where \( n_d \) is the number of dead objects.

The addition and removal actions change the dimension of the state vector. When calculating the acceptance probability, we need to compute the ratio of probabilities from spaces with different dimensions. Smith et al. [41] use an explicit strategy of trans-dimensional MCMC [14] to deal with the dimension-matching problem. We do not need explicit strategy to match the dimension. Since the trans-dimensional actions only add or remove one object at one iteration, leaving the other objects unchanged, the Jacobian in [14] is unit, as in [41]. So our formulation is just a special case of the more general theory.

**Parameter update** We use two ways to update the model parameters: \( q_{d_{1f}}(m_i'|m_r) = \lambda_{d_1} q_{d_1} (m_i'|m_r) + \lambda_{d_2} q_{d_2} (m_i'|m_r) \), \( \lambda_{d_1} = 1/2. q_{d_1}() \) uses stochastic gradient decent to update the object parameters. \( q_{d_1} (m_i'|m_r) \propto N\left(\bar{m}_r - k \Delta d_{im}, w\right) \), where \( E = -\log P(\theta(t)|I(t), \theta(t-1)) \) is the energy function, \( k \) is a scalar to control the step size, and \( w \) is random noise to avoid local maximum.

A mean shift vector computed in the visible region provides an approximation of the gradient of the object likelihood w.r.t. the position. \( q_{d_{2f}}(m_i'|m_r) \propto N(m_i''|m_r) \), where \( m_i'' \) is the new location computed from the mean shift procedure (details are given in a separate Appendix). We assume that the change of the posterior probability by other components and due to occlusion can be absorbed in the noise term. The mean shift has an adaptive step size and has a better convergence behavior than numerically computed gradients. The rest of the parameters follow their numerically computed gradients. Compared to the original color-based mean shift tracking, the background exclusion term in Equation 6, can utilize a known background model, which is available for a stationary camera. As we observe in our experiments, tracking using the above likelihood is more robust to the change of appearance of the object, e.g. when going into the shadow, compared to using the object attraction term alone.

Theoretically, the Markov chain designed should be irreducible and reversible, however the use of the above data driven proposal probabilities makes the approach not conform to the theory exactly. First, irreducibility requires the Markov chain be able to reach any possible point in the solution space. However, in practice, the proposal probability of some point are very small, close to zero. For example the proposal probability of adding a hypothesis at a position, where there is no head candidate detected nearby, is extremely low. With finite numbers of iterations, a state including such a hypothesis will never be sampled. Although this breaks the completeness of the Markov chain, we argue that skipping the parts of the solution space, where no sign of objects observed, brings no harm to the quality of the final solution and makes the searching process more efficient. Second, the use of the mean shift, which is a non-parametric method, makes the chain irreversible. Mean-shift can be seen as an approximation of the gradient, while stochastic gradient decent is essentially a Gibbs sampler [39], which is a special case of Metropolis-Hasting sampler with acceptance ratio always equal to one [25]. However, mean shift is much faster than the random walk to estimate the parameters of the object. We choose to use these techniques with the lost of some theoretical beauty, because experimentally they makes our method much more efficient and the results are good.

### C. Incremental Computation

As the MCMC process may need hundreds or more samples to approximate the distribution, we need an efficient method to compute the likelihood for each proposed state. In one iteration of the algorithm, at most two objects may change. It effects the likelihood locally, therefore the computation of the new likelihood can be carried out more efficiently by incrementally computing it only within their neighborhood (the area associated with the changed objects and those overlapping with them).

Take the addition action as an example. When a new human hypothesis is added to the state vector, for the likelihood of the non-object region \( P(I^5|\theta) \), we only need to remove those background pixels taken by the new hypothesis. For the likelihood of the object region \( P(I^5|\theta) \), as the new hypothesis may overlap with some existing hypotheses, we need to recompute the visibility of the object regions connected to the new hypothesis and then update the likelihood of these neighboring objects. The incremental computations of the likelihood for the other actions are similar. Although a joint state and joint likelihood is used, the computation of each iteration is greatly reduced through the incremental computation. This is in contrast to the particle filter where the evaluation of each particle (joint state) needs the computation of the full joint likelihood.

The appearance models of the tracked objects are updated after processing each frame to adapt to the change in object appearance. We update the object color histogram using an IIR filter \( \tilde{p}^{(t)} = \lambda_p p^{(t)} + (1 - \lambda_p)\tilde{p}^{(t-1)} \). We choose to update the appearance conservatively: we use a small \( \lambda_p = 0.01 \) and stop updating if the object is occluded by more than 25% or its position covariance is too big.

### VI. Experimental Results

We have experimented the system with many types of data and will only show some representative ones. We will first show results on an outdoor scene video and then on a standard evaluation dataset of indoor scene videos. Video results are submitted as supplementary materials.

Among all the parameters of our approach, many are “natural”, meaning that they correspond to measurable physical quantities (e.g. 3d human height), therefore setting their values is straightforward. We use the same set of parameters for all the sequences. This means that our approach is not sensitive to the choice of parameter values. We list here the values of the parameters which are not mentioned in the previous sections. For the size prior (in Sec. IV-D), \( \lambda_1 = 0.04 \) and \( \lambda_2 = 0.002 \). For likelihood, \( \lambda_1 = 0.5, \lambda_2 = 0.5 \) in Equation 6, \( \lambda_S = 25 \) in Eqn. 6 and \( \lambda_{\theta} = 0.005 \) in Eqn. 7. For the mixing probabilities of different types of dynamics, we use...
We also want to comment here on the choice of parameters related to the peakedness of a distribution in sampling algorithms. The image likelihood is usually a combination of a number of components (sites, e.g. pixels). Inevitable simplifications (e.g. independence assumption) in probabilistic modeling may result in excessive peakedness of the distribution, which affects the performance of the sampling algorithms such as MCMC and particle filter by having the samples in both MCMC and particle filter focused in one location (i.e. highest peak) of the state space therefore makes them to degenerate into greedy algorithms. Eliminating the dependencies of different components can be extremely difficult and infeasible. From an engineering point of view, one should set the values of the parameters (e.g. \( \lambda_S \) and \( \lambda_\Sigma \) while keeping their ratio constant) so that likelihood ratio of different hypotheses are reasonable, so that the Markov chains can efficiently traverse and particle filters can maintain multiple hypotheses. In a similar fashion, simulated annealing has been used in the sampling process to reduce the effect of the peakedness and force convergence \([48], [8]\), however the varying temperature makes the samples not from a single posterior distribution.

A. Evaluation on an Outdoor Scene

We show results on an outdoor video sequence, that we call the “Campus Plaza” sequence, which contains 900 frames. This sequence is captured from a camera above a building gate with a 40° camera tilt angle. The frame size is 360 × 240 pixel, and the sampling rate is 30 FPS. In this sequence, 33 humans pass by the scene with 23 going out of field of view and 10 going inside a building. The inter-human occlusions in this sequence are large. There are overall 20 occlusion events, 9 out of them are heavy occlusion (over 50% of the object is occluded). For MCMC sampling, we use 500 iterations per frame. We show in Fig.7 some sample frames from the result on this sequence. The identities of the objects are shown by their ID numbers displayed on the head.

We evaluate the results by the trajectory-based errors. Trajectories whose lengths are less than 10 frames are discarded. Among the 33 human objects, trajectories of 3 objects are broken once (ID 28→ID 35, ID 31→ID 32, ID 30→ID 41, all between frame 387 and frame 447, as marked with arrows in Fig.7); rest of the trajectories are correct. Usually the trajectories are initialized once the humans are fully in the scene, some start when the objects are only partially inside. Only the initializations of three objects (objects 31, 50, 52) are noticeably delayed (by 50, 55, 60 frames respectively after they are fully in the scene). Partial occlusion or lack of contrast with the background are the causes of the delays. To justify our approach for integrated segmentation and tracking, we compare the tracking result with the result using frame-by-frame segmentation as in \([53]\) where we use frame-based evaluation metrics. The detection rate and the false alarm rate of the same sequence by using segmentation alone are 92.82% and 0.18%. With tracking, not only the temporal correspondences are obtained, but also the detection rate is increased by a large margin while the false alarm rate is kept low.

B. Evaluation on Indoor Scene Sequences

We evaluate the results by the trajectory-based errors. Trajectories whose lengths are less than 10 frames are discarded. Among the 33 human objects, trajectories of 3 objects are broken once (ID 28→ID 35, ID 31→ID 32, ID 30→ID 41, all between frame 387 and frame 447, as marked with arrows in Fig.7); rest of the trajectories are correct. Usually the trajectories are initialized once the humans are fully in the scene, some start when the objects are only partially inside. Only the initializations of three objects (objects 31, 50, 52) are noticeably delayed (by 50, 55, 60 frames respectively after they are fully in the scene). Partial occlusion or lack of contrast with the background are the causes of the delays. To justify our approach for integrated segmentation and tracking, we compare the tracking result with the result using frame-by-frame segmentation as in \([53]\) where we use frame-based evaluation metrics. The detection rate and the false alarm rate of the same sequence by using segmentation alone are 92.82% and 0.18%. With tracking, not only the temporal correspondences are obtained, but also the detection rate is increased by a large margin while the false alarm rate is kept low.

Next, we describe the results of our method on an indoor video set, CAVIAR video corpus\(^4\) \([56]\). We test our system on the 26 “shopping center corridor view” sequences, overall 36,292 frames, captured by a camera looking down towards a corridor. The frame size is 384 × 288 pixel, and the sampling rate is 25 FPS. Some 2D-3D point correspondences are given from which the camera can be calibrated. However, we compute the camera parameters by an interactive method \([26]\).

The inter-object occlusion in this set is also intensive. There are overall 96 occlusion events in this set, 68 out of 96 are heavy occlusions, and 19 out of the 96 are almost fully occlusions (more than 90% of the object is occluded). Many interactions between humans, such as talking, and hand shaking, make this set very difficult for tracking. For MCMC sampling, we use 500 iterations per frame. For such a big data set, it’s infeasible to enumerate the errors like for the “Campus Plaza” sequence. Instead we defined five statistical criteria: 1) number of mostly tracked trajectories; 2) number of mostly lost trajectories; 3) number of fragments of trajectory; 4) number of false trajectories (a results trajectory corresponding to no object); and 5) the frequency of identity switches (identity exchanging between a pair of result trajectories). Fig.8 illustrates their definition. These five categories are by no means a complete classification, however they cover most of the typical errors observed on this set. There are other performance measures that have been proposed in the recent evaluations, such as the Multiple Object Tracking Precision and Accuracy in the CLEAR 2006 evaluation \([57]\). We do not use these measures, because they are less intuitive, as they try to integrate multiple factors into one scalar valued measure.

Table I gives the performance of our method. We developed an evaluation software to count the number of mostly tracked trajectories, mostly lost trajectories, false alarms and fragments automatically. Denote a ground-truth trajectory by \(\{G^{(i)}\}_{i=1}^{l} \), where \(G^{(i)}\) is the object state at the \(t\)-th frame; denote a hypothesized trajectory by \(\{H^{(j)}\}_{j=1}^{m} \). The overlap ratio of the ground-truth

\(^{4}\)In the provided ground-truth, there are 232 trajectories overall. However 5 of these are mostly out of sight, e.g. only one arm or the head top is visible; we set these as “don’t care”.

![Fig. 8. Tracking evaluation criteria.](image-url)
Fig. 7. Selected frames of the tracking results from “Campus Plaza”. The numbers on the heads show identities. (Please note that the two people who are sitting on two sides are in the background model, therefore not detected.)
object and the hypothesized object at the \( t \)-frame is defined by

\[
\text{Overlap}(\mathbf{G}(t), \mathbf{H}(t)) = \frac{\text{Reg}(\mathbf{G}(t)) \cap \text{Reg}(\mathbf{H}(t))}{\text{Reg}(\mathbf{G}(t)) \cup \text{Reg}(\mathbf{H}(t))}
\]  

(8)

where \( \text{Reg}(\cdot) \) is the image region of the object. If \( \text{Overlap}(\mathbf{G}(t), \mathbf{H}(t)) > 0.5 \), we say \( \{\mathbf{G}(t), \mathbf{H}(t)\} \) is a potential match. The overlap ratio of the ground-truth trajectory and the hypothesized trajectory is defined by

\[
\text{Overlap}(\mathbf{G}(t), \mathbf{H}(t)) = \frac{\sum_{t_{\min}(i+n,j+m)}^{t_{\max}(i+n,j+m)} \delta(\text{Overlap}(\mathbf{G}(t), \mathbf{H}(t)) > 0.5)}{\max(i+n,j+m)-\min(i,j)+1}
\]  

(9)

where \( \delta(\cdot) \) is an indicator function. Given that one sequence has \( N_G \) ground-truth trajectories \( \{\mathbf{G}_k\}_{k=1}^{N_G} \), and \( N_H \) hypothesized trajectories \( \{\mathbf{H}_k\}_{k=1}^{N_H} \), we compute the overlap ratios for all ground-truth hypothesis pairs \( \{\mathbf{G}_k, \mathbf{H}_j\} \); the pairs whose overlap ratios are larger than 0.8 are considered to be potential matches. Then the Hungarian matching algorithm [22] is used to find the best matches which are considered to be mostly tracked. To count the mostly lost trajectories, we define a recall ratio by replacing the denominator of Eq.(9) with \( n + 1 \). If for \( \mathbf{G}_k \), there is no \( \mathbf{H}_j \) such that the recall ratio between them is larger than 0.2, we consider \( \mathbf{G}_k \) to be mostly lost. To count the false alarm and fragments, we define a precision ratio by replacing the denominator of Eq.(9) with \( m + 1 \). If for \( \mathbf{H}_j \) there is no \( \mathbf{G}_k \) such that the precision ratio between them is larger than 0.2, we consider \( \mathbf{H}_j \) a false alarm; if there is such a \( \mathbf{G}_k \) that the precision between them is larger than 0.8, but the overlap ratio is smaller than 0.8, we consider \( \mathbf{H}_j \) to be a fragment of \( \mathbf{G}_k \). We first count the mostly tracked trajectories, and remove the matched parts of the ground-truth tracks. Second, we count the trajectory fragments with a greedy, iterative algorithm. At each round, the fragment with the highest overlap ratio is found, and then the matched part of the ground-truth track is removed; this procedure is repeated until there are no more valid fragments. Lastly, we count the mostly lost trajectories and the false alarms. This algorithm can not classify all ground-truth and hypothesized tracks; the unlabeled ones are mainly due to an identity switch. We count the frequency of identity switches visually.

Some sample frames and results are shown in Fig.9. Most of the missed detections are due to the humans wearing clothing with color very similar to that of the background so that some part of the object is misclassified as background, see the frame 1413 of Fig.9(b) for an example. The fragmentation of trajectory and the ID switch are mainly due to full occlusions, see the frame 496 of Fig.9(a) and the frame 316 of Fig.9(b) for examples. Our method can deal with partial occlusion well. For full occlusion, classifying an object as going into an “occluded” state and associating it when it reappears could potentially improve the performance. The false alarms are mainly due to the shadows, reflections and sudden brightness changes which are misclassified as foreground, see the frame 563 of Fig.9(a). More sophisticated background model and shadow model (e.g. [32]) could be used to improve the result. In general, our method performs reasonably well on the CAVIAR set, though not as well as on the “Campus Plaza” sequence, mainly due to the above mentioned difficulties. The running speed of the system is about 2 FPS with a 2.8GHz Pentium IV CPU. The implementation is in C++ code without any special optimization.

### VII. Conclusion and Future Work

We have presented a principled approach to simultaneously detect and track humans in a crowded scene acquired from a single stationary camera. We take a model-based approach and formulate the problem as a Bayesian MAP estimation problem to compute the best interpretation of the image observations collectively by the 3D human shape model, acquired human appearance model, background appearance model, camera model, the assumption that humans move on a a known ground plane, and the object priors. The image is modeled as a composition of an unknown number of possibly overlapping objects and a background. The inference is performed by an MCMC-based approach to explore the joint solution space. Data-driven proposal probabilities are used to direct the Markov chain dynamics. Experiments and evaluations on challenging real-life data show promising results.

The success of our approach mainly lies in the integration of the top-down Bayesian formulation following the image formation process and the bottom-up features that are directly extracted from images. The integration has the benefit of both the computational efficiency of image features and the optimality of a Bayesian formulation.

This work could be improved/extended in several ways. 1) extension to track multiple classes of objects (e.g. humans and cars), by adding model switching in the MCMC dynamics. 2) Tracking, operating in a 2-frame interval, has a very local view therefore ambiguities inevitably exist, especially in the case of tracking fully occluded objects. The analysis in the level of trajectories may resolve the local ambiguities (e.g. [29]). The analysis may take into account the prior knowledge on the valid object trajectories including their starting and ending points.

### APPENDIX I

**Single Object Tracking with Background Knowledge Using Meansift**

Denote by \( p_\mathbf{p}(\mathbf{u}) \), \( p_\mathbf{b}(\mathbf{u}) \), and \( b(\mathbf{u}) \) the color histograms of the object learnt online, the color histogram of the object at location \( \mathbf{u} \) and the color histogram of the background at the corresponding region respectively. Let \( \{x_i\}_{i=1,\ldots,n} \) be the pixel locations in the region with the object center at \( \mathbf{u} \). A kernel with profile \( k(\cdot) \) is used to assign smaller weights to the pixels farther away from the center. An \( m \)-bin color histogram \( p_\mathbf{p}(\mathbf{u}) = \{p_j(\mathbf{u})\}_{j=1,\ldots,m} \) is constructed as \( p_j(\mathbf{u}) = \)
\[
\sum_{i=1}^{n} k \left( \|x_i\|^2 \right) \delta [b_f (x_i) - j], \quad \text{where function } b_f() \text{ maps the}\]
\[
\text{pixel location to the corresponding histogram bin, and } \delta \text{ is the}\]
\[
delta function. Similar for } \tilde{p} \text{ and } b. \text{ We would like to optimize}\]
\[
L(u) = -\lambda_b B(p(u), b(u)) + \lambda_f B(p(u), \tilde{p})
\]
\[\frac{L_1(u)}{L_2(u)}\]
\[
(10)
\]
\[
\text{where } B() \text{ is the Bhattachayya coefficient. By applying Taylor}\]
\[
\text{expansion at } \mathbf{p}(\mathbf{u}_0) \text{ and } \mathbf{b}(\mathbf{u}_0) \text{(a predicted position of}\]
\[
\text{the object), we have}\]
\[
L_1(u) = B(p(u), b(u)) = B(u)
\]
\[
\approx B(u_0) + B'_p(u_0) \mathbf{p}(\mathbf{u}) - \mathbf{p}(\mathbf{u}_0)) + B'_b(u_0) \mathbf{b}(\mathbf{u}) - \mathbf{b}(\mathbf{u}_0))
\]
\[
= c_1 + \sum_{i=1}^{m} \frac{b_i(\mathbf{u}_0)}{p_i(\mathbf{u}_0)} p_i(\mathbf{u}) + \sum_{i=1}^{m} \frac{p_i(\mathbf{u}_0)}{b_i(\mathbf{u}_0)} b_i(\mathbf{u})
\]
\[
= c_1 + \sum_{i=1}^{m} k \left( \|u - x_i\| / h \right) w_i^b
\]
\[\tag{11}
\]
\[
\text{where}\]
\[
w_i^b = \sum_{u=1}^{m} \left\{ \frac{b_i(\mathbf{u}_0)}{p_i(\mathbf{u}_0)} \delta [b_f (x_i) - u] + \sqrt{\frac{p_i(\mathbf{u}_0)}{b_i(\mathbf{u}_0)}} \delta [b_b (x_i) - u] \right\}
\]
\[
\text{Similarly, also in } [6],\]
\[
L_2(u) = B(p(u), \tilde{p}) \approx \frac{1}{2} \sum_{i=1}^{m} p_i(\mathbf{u}_0) p_i(\mathbf{u}) + \frac{1}{2} \sum_{i=1}^{m} p_i(\mathbf{u}_0) \sqrt{\frac{\tilde{p}_i(\mathbf{u}_0)}{p_i(\mathbf{u}_0)}}
\]
\[
= c_2 + \sum_{i=1}^{m} w_i^f k \left( \|u - x_i\| / h \right)^2
\]
\[\tag{12}
\]
\[
\text{for}\]
\[
\text{The last term of } L(u) \text{ is the density estimate computed with}\]
\[
\text{kernel profile } k() \text{ at } u. \text{ The meanshift algorithm with negative}\]
weight [4] applies. By using the Epanechnikov profile [6], \( L(u) \) will be increased with the new location moved to

\[
\mathbf{u}' \leftarrow \frac{\sum_{i=1}^{n} x_i u_{i|u_i}}{\sum_{i=1}^{n} |w_i|}
\]

(14)

ACKNOWLEDGMENT

This research was funded, in part, by the U.S. Government VACE program.

REFERENCES


Tao Zhao received the BEng degree from the Department of Computer Science and Technology, Tsinghua University, China, in 1998. He received the MSc and the PhD degrees from the Department of Computer Science at the University of Southern California in 2001 and 2003, respectively. He was with Sarnoff Corporation, Princeton, New Jersey, from 2003 to 2006. He is currently with Intuitive Surgical Incorporated, Sunnyvale, California working on computer vision applications for medicine and surgery. His research interests include computer vision, machine learning, and pattern recognition. His experience has been in visual surveillance, human motion analysis, aerial image analysis and medical image analysis. He is a member of the IEEE and IEEE computer society.

Ram Nevatia received his Ph.D. degree from Stanford University with specialty in the area of computer vision. He has been with the University of Southern California since 1975 where he is currently a Professor of Computer Science and Electrical Engineering. He is also Director of the Institute for Robotics and Intelligent Systems. He has been principal investigator of major Government funded computer vision research programs for over 25 years. Dr. Nevatia has made important contributions to several areas of computer vision including the topics of shape description, object recognition, stereo analysis aerial image analysis, tracking of humans and event recognition. Dr. Nevatia is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE) and of the American Association for Artificial Intelligence (AAAI). He is an associate editor for the Pattern Recognition, and the Computer Vision and Image Understanding journals. Dr. Nevatia is author of two books, several book chapters, and over 100 refereed technical papers.

Bo Wu received the B.Eng and M.Eng degrees from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2002 and 2004 respectively. He is currently a PhD candidate at the Computer Science Department, University of Southern California, Los Angeles. His research interests include computer vision, machine learning, and pattern recognition. He is a student member of the IEEE computer society.