Bayesian Human Segmentation in Crowded Situations *

Tao Zhao Ram Nevatia
University of Southern California
Institute for Robotics and Intelligent Systems
Los Angeles, CA 90089-0273
{taozhao|nevatia}@usc.edu

Abstract

Problem of segmenting individual humans in crowded situations from stationary video camera sequences is exacerbated by object inter-occlusion. We pose this problem as a "model-based segmentation" problem in which human shape models are used to interpret the foreground in a Bayesian framework. The solution is obtained by using an efficient Markov chain Monte Carlo (MCMC) method which uses domain knowledge as proposal probabilities. Knowledge of various aspects including human shape, human height, camera model, and image cues including human head candidates, foreground/background separation are integrated in one theoretically sound framework. We show promising results and evaluations on some challenging data.

1 Introduction and Motivation

Segmentation and tracking of humans in video sequences are important for a number of tasks such as video surveillance and event inference as humans are the principal actors in daily activities of interest. In the situation of stationary cameras, background subtraction is a widely used technique to extract the moving pixels (foreground). If objects are sparse in the scene, each connected component of the foreground (*blob*) usually corresponds to an object, though the blobs may be fragmented due to low contrast or partial occlusion by scene objects or may contain non-human pixels caused by shadow. When the density of the objects increases (such as in the example shown in Fig.1), it is common that several objects form one big blob. Therefore the blobs do not directly provide the object level description that is needed for human event inference. Our goal in this paper is to segment the foreground (a binary mask) into individual human objects which may overlap with each other.

1.1 Previous work

Some work has been done to segment or track multiple overlapping humans. In [5], peaks in the vertical histogram of the blob are used to help locate the positions of the heads. In [16] and [11], vertical peaks on the foreground boundary are used to locate the positions of the heads. [16] also employs an iterative processing to handle the case when the heads are not on the foreground boundary, assuming that the occlusion from each other is slight. In [3], humans are assumed to be isolated as they enter the scene so that a human specific color model can be initialized for segmentation when occlusion occurs. [16] and [11] also use the initialized human specific model to help in tracking through occlusion. All of the above techniques either are based on some heuristics for segmentation or rely on initialized human models before occlusion occurs. They are not likely to be effective in crowded situations.

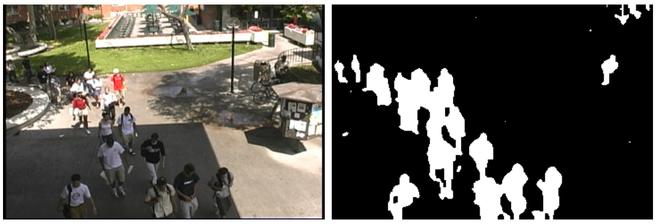

(a) (b)

Figure 1: A sample input frame (a) and its foreground from standard background subtraction (b).

In [10], the segmentation problem is solved using region-based stereo with input from multiple (up to 16) cameras. However it is limited to applications of small spatial areas.

[12] and [8] propose work to track multiple humans using *particle filters*. Due to the possible inter-occlusions, joint states (the parameters of all the humans in the scene) are used in the tracking. Performance of particle filters is limited by the dimentionality of the state space, therefore, extending these approaches to track a large number of humans may be difficult.

Color segmentation is not likely to segment individual humans. Motion segmentation also may not give satisfactory result due to the non-rigid human motion and the similarity of the motion of individuals in a group. Face detection may not be effective when the humans don't face the camera or when the image size of the human is small. Direct human detection (*e.g.*, [9]) has been limited to restricted viewpoints (frontal or back).

[1] This research was supported, in part, by the Advanced Research and Development Activity of the U.S. Government under contract No. MDA-908-00-C-0036.

1.2 Our approach

We take a 3D model-based approach and use human shape models to interpret the foreground. The problem is defined as a *model-based segmentation* problem in a Bayesian framework. The solution is defined to be the number of human objects and their associated parameters maximizing the posterior probability. The posterior probability reflects how well the solution reconstructs the foreground by the image likelihood while preferring small number of objects by the prior.

We use MCMC with jump and diffusion dynamics to pursue the best solution by traversing the solution space. Markov chain Monte Carlo (MCMC) is a tool to sample a probabilistic distribution. Recently DDMCMC (data-driven MCMC) has been proposed to solve computer vision problems. It improves the efficiency of traditional MCMC by incorporating domain knowledge to compute proposal probabilities of the Markov chain especially when the state is defined on a complex space. Promising results have been shown on object recognition [18], image segmentation [14] and range image segmentation [7].

Domain knowledge including head candidates computed from foreground boundaries, head candidates computed from intensity edges and analysis of foreground residue map direct the creation of new human hypotheses which improves the efficiency of the MCMC significantly.

The described work builds on our earlier work reported in [17], The differences include more expressive human models to capture the human shape variations in mid-range data (*i.e.* where major limb articulations can be seen, stochastic diffusion which both improves the speed and the accuracy of localization and other refinements that result in more robust performance.

We have performed experiments on datasets of different object densities and under different imaging conditions. The proposed approach gives good results with affordable computation time as described later in the paper.

2 A Bayesian Formulation of the Problem

We formulate the segmentation problem as computing the maximum a posteriori (MAP) estimation θ^* such that

$$\theta^* = argmax_{\theta \in \Theta} P(\theta | I) \qquad (1)$$

where θ is the number of human objects and their parameters and I is the foreground mask. Following Bayes rule, the posterior probability is decomposed into a likelihood term and a prior term:

$$P(\theta|I) \propto P(I|\theta)P(\theta) \qquad (2)$$

The human shape model, the prior and the likelihood model are described in this section.

2.1 3D human shape model

Human body shape is highly articulated. To model it precisely, a kinematics model with over 20 DOF and a mass distribution model are needed. However, in our application the human motion is mostly limited to standing or walking and we do not attempt to capture the detailed parameters of the human body. We can thus use a number of low dimensional models to capture the gross shape of human bodies at different pose/viewpoint combinations.

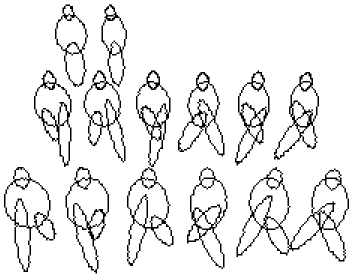


Figure 2: A number of 3D human models are used to capture the gross shape of standing and walking humans. First row: standing models (with orientation 0^o and 90^o); second/third row: walking models with left/right leg forward (with orientation 0^o, $\pm30^o$, $\pm60^o$ and 90^o). NOTE: they are 2D projection of 3D models under the camera model of Seq.2 in the result section.

We model human shape by four ellipsoids corresponding to head, torso and two legs. An ellipsoid fits human body parts well and has the property that its projection is an ellipse with a convenient form [6]. Each ellipsoid is controlled by two parameters called *length* and *fatness*. The length parameter also determines the width by using a fixed ratio; fatness parameter defines depth besides the proportional change. We consider only three articulations: both legs together, left leg forward and right leg forward. These models are sufficient to capture the gross shape variations of most humans in the scene for mid-resolution images.

We assume that the humans move on a ground plane, therefore, besides *height* and *fatness*, the parameters of the model also include *position* on the ground plane and *orientation*. The orientations of the models are quantized for computation efficiency as follows: the standing model has two orientations (0^o, frontal and 90^o, from side) and each of the two walking models has six orientations (0^o, $\pm30^o$, $\pm60^o$ and 90^o). Since both the model type and the orientation are discrete, we combine them into one parameter *model/orientation label* in one of the 14 combinations (Fig.2). Therefore, the parameters of each human object i are $M_i = \{l_i, x_i, y_i, h_i, f_i\}$ which are *model/orientation label*, *position*, *height* and *fatness* respectively.

We assume that the camera model and the ground plane are known as in [16]. The camera model and the 3D shape model automatically take care of the change in 2D size and shape due to the change in position and viewpoint. This is advantageous to 2D shape models such as in [4].

2.2 The solution space

The solution to the model-based segmentation problem includes the number of objects in the scene and their associated parameters. It can be written in the form of $\theta = \{n, \{M_1, M_2, ..., M_n\}\} \in \Theta$.

$$\Theta = \cup_{n=0}^{\infty}\Theta_n, \Theta_n = (L \times R^4)^n$$

where Θ_n is the subspace of exactly n objects, L is the enumerated space of the model/orientation ($|L| = 14$) and R^4 is the space for position and shape parameters (m_i).

Since we don't know in advance how many objects are present in a given scene, the solution space Θ contains subspaces of varying dimensions.

2.3 Prior distributions

We assume the prior probability of the state is the product of the prior probabilities of all the objects. The prior probability of an object i is made up of the prior probability on its image size (A_i) and the prior probability on its parameters (M_i).

$$P(\theta) = \Pi_{i=1}^{n} P(A_i)P(M_i) \tag{3}$$

$$P(A_i) = e^{-\lambda_1 A_i}(1 - e^{-\lambda_2 A_i})$$

The first term penalizes large total object sizes which avoids unnecessary overlapping. The second term penalizes objects with small image sizes since they are more likely to be due to image noise ($e^{-\lambda_2 A_i}$ is the distribution of the noise blob size). λ_1 is a coefficient which controls the maximum overlap allowed in the final segmentation and λ_2 is related to the noise level of the images. We also experimented with prior probabilities of the number of objects ($P(n) = e^{-\lambda_1' n}$) but found it to be not effective in scenes which contain human objects with large image size variations.

$$P(M_i) = P(l_i)P(x_i, y_i)P(h_i)P(f_i)$$

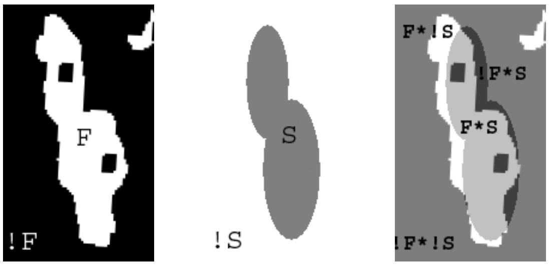
We set $P(l_i)$ so that $P(l_{walk}) = 2/3P(l_{stand})$ to penalize the complexity (more ellipsoid) of the walking models. $P(x_i, y_i)$ is a uniform distribution in the image. $P(h_i)$ is a Gaussian distribution $N(\mu_h, \sigma_h^2)$ ($\mu_h = 1.7m$, $\sigma_h = 0.2$) truncated in the range of $[1.5m, 1.9m]$ and $P(f_i)$ is Gaussian distribution $N(\mu_f, \sigma_f^2)$ ($\mu_f = 1$, $\sigma_f = 0.2$) truncated in the range of $[0.8, 1.2]$. Therefore:

$$P(M_i) = P(l_i)e^{-(\frac{h_i - \mu_h}{\sigma_h})^2} e^{-(\frac{f_i - \mu_f}{\sigma_f})^2};$$

$$h_i \in [1.5, 1.9], f_i \in [0.8, 1.2]$$

2.4 Multi-object joint likelihood

Since multiple humans may occlude each other, the image likelihood cannot be decomposed into the product of image likelihoods individual human hypotheses. Given a state, we compute joint likelihood based on the formation of the foreground as follows.



(a) (b) (c)

Figure 3: The likelihood based on the number of wrongly classified pixels. (a) the foreground F; (b) the region of the solution S; (c) four kinds of pixels. NOTE: ellipse model is used for illustration.

We assume that the foreground is formed by human objects in the scene. Denote F as the image foreground and S as the union of image regions of all human objects in a solution θ (see Fig.3 for F, S and their relationship). p_{11} is the probability that a pixel in an human object results in a foreground pixel, p_{10} is the probability that a pixel in an human object does not result in a foreground pixel, p_{01} is the probability that a pixel outside human objects results in a foreground pixel, p_{00} is the probability that a pixel outside human objects does not create a foreground pixel. $p_{11} + p_{10} = 1, p_{11} > p_{10}$ and $p_{01} + p_{00} = 1, p_{01} < p_{00}$. Assuming the pixels are independent, we have the following likelihood:

$$P(I|\theta) = \Pi_{i \in I} P(i|\theta) =$$

$$\Pi_{i \in F*S} P(i|\theta) \Pi_{i \in F*\overline{S}} P(i|\theta) \Pi_{i \in \overline{F}*S} P(i|\theta) \Pi_{i \in \overline{F}*\overline{S}} P(i|\theta)$$

$$= p_{11}^{\Sigma_{i \in F*S}} p_{10}^{\Sigma_{i \in F*\overline{S}}} p_{01}^{\Sigma_{i \in \overline{F}*S}} p_{00}^{\Sigma_{i \in \overline{F}*\overline{S}}}$$

$$= p_{11}^{N_{11}} p_{10}^{N_{10}} p_{01}^{N_{01}} p_{00}^{N_{00}}$$

$$= (1 - p_{01})^{|F| - N_{01}} p_{10}^{N_{10}} p_{01}^{N_{01}} (1 - p_{10})^{|\overline{F}| - N_{10}}$$

$$= (1 - p_{01})^{|F|} (1 - p_{10})^{|\overline{F}|} (\frac{p_{10}}{1 - p_{10}})^{N_{10}} (\frac{p_{01}}{1 - p_{01}})^{N_{01}}$$

$$= \alpha e^{-(\lambda_{10} N_{10} + \lambda_{01} N_{01})}$$

where α is a constant which absorbs the terms independent of θ, $\lambda_{10}, \lambda_{01}$ are two coefficients depending on p_{10} and p_{01} and $*$ is the intersection of two regions. The likelihood only depends on N_{10} and N_{01}, which are the difference of the foreground F and the solution region S. The parameters p_{01} and p_{10} are estimated by examples.

By combining the prior (Equ.3) and the above likelihood, the posterior probability (Equ.2) becomes:

$$P(\theta|I) \propto$$

$$(\Pi_{i=0}^{n} e^{-\lambda_1 A_i}(1 - e^{-\lambda_2 A_i}) P(l_i) e^{-(\frac{h_i - \mu_h}{\sigma_h})^2} e^{-(\frac{f_i - \mu_f}{\sigma_f})^2}) \cdot$$

$$e^{-(\lambda_{10} N_{10} + \lambda_{01} N_{01})}; h_i \in [1.5, 1.9], f_i \in [0.8, 1.2]$$

3 Efficient MAP Computation

We want to find the segmentation that maximizes the posterior probability defined in the previous section. However, as stated in Sec.2.2, the solution space contains subspaces of varying dimensions and may contain many local minimums. Markov chain Monte Carlo (MCMC) methods combined with jump-diffusion dynamics provides a way to sample the posterior probability in such a complex solution space to search for the maximum.

The basic idea of MCMC is as follows. A Markov chain can be designed to sample a probability distribution $Q(\theta)$ (we use $Q(\theta) = P(\theta|I)$ for clarity). At each iteration t, we sample a candidate state θ' according to θ_{t-1} from a *proposal distribution* $q(\theta_t|\theta_{t-1})$ (in simple words, what new state should the Markov chain go to from the previous state.). The candidate state θ' is accepted with the following probability:

$$p = min(1, \frac{Q(\theta')q(\theta_{t-1}|\theta')}{Q(\theta_{t-1})q(\theta'|\theta_{t-1})})$$

If the candidate state θ' is accepted, $\theta_t = \theta'$, otherwise, $\theta_t = \theta_{t-1}$. This is the well-known Metropolis-Hasting algorithm. It can be proven that the Markov chain constructed this way has its stationary distribution equal to $Q()$, independent of the choice of the proposal probability $q()$ and the initial state θ_0 [13]. However, the choice of the proposal probability $q()$ can affect the efficiency of the MCMC significantly. A random proposal probability will lead to very slow convergence rate while a proposal probability designed with domain knowledge ([18] [14] [7]) will make the Markov chain traverse the solution space more efficiently. If the proposal probability is informative enough so that each sample can be thought of as a *hypothesis*, then the MCMC approach can be thought of as a stochastic version of the *hypothesize and test* approach ([18]). Sampling the proposal probability results in certain dynamics of the Markov chain. Jump means that the structure (*e.g.* dimension) of the state is change while diffusion means that the structure does not change but the values of the parameters change.

The crucial point in this problem is where to add a human object. We incorporate three types of domain knowledge in proposing the positions of new human objects.

3.1 Hypothesizing human objects

3.1.1 Head candidates from foreground boundary

This method detects the heads which are on the boundary of the foreground [16]. The basic idea is to find the local vertical peaks of the boundary. The peaks are further filtered by checking if there are enough foreground pixels below it according to the 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.4.(a) shows the $HC_{boundary}$ on the example frame.

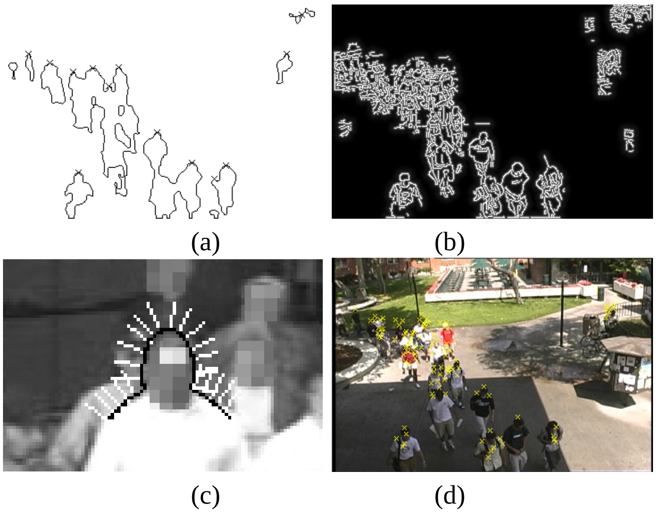


Figure 4: Head detectors. (a) Head candidates (crosses) from foreground boundaries ($HC_{boundary}$); (b) Distance transformation on Canny edge detection result; (c) The head-shoulder (Ω) model: dark contour-head and shoulder, light line-normals; (d) Head candidates (crosses) from intensity edges (HC_{edge}). Read Sec.3.1 for detail.

3.1.2 Head candidates from intensity

we also use a head detector based on image intensity edges which is also effective for the heads in the interior of the blobs [17]. First, a Canny edge detector [2] is applied to the dilated foreground region of the input image. A distance transformation [1] is then computed on the edge map. Fig.4.(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 λ is a factor to control the response field and set to 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 of head and shoulder contour (Fig.4.(c)) is easily derived from our human model. The head-shoulder contour is generated from the projected ellipses by taking 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 a known height (1.6, 1.7, 1.8 meters are used and the maximum response is recorded).

Denote $\{\vec{m}_1, ..., \vec{m}_k\}$ 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 in the following way.

$$S(x,y) = (1/k)\Sigma_{i=1}^{k} e^{-\lambda D(\vec{m}_i)}(\vec{v}_i \cdot \vec{O}(\vec{C}(\vec{m}_i)))$$

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 selected to give a very high detection rate but may also result in a high false alarm rate. An example is shown in Fig.4.(d). The false alarms tend to happen in the areas of rich texture where there are abundant edges of various orientations.

3.1.3 Residue analysis

We denote the foreground map with the already formed hypotheses removed as the foreground residue map. The foreground may initially contain some blobs of almost separated human objects. After some human objects are hypothesized and removed from the foreground, the residue map may show more separated objects. Morphological *open* operation can help isolate the objects that are only connected by thin bridges and remove small/thin residues. We generate human candidates from the foreground residue map as follows.

Given foreground F and the region S corresponding to the current solution θ, compute the foreground residue map $R = F * \overline{S}$. Perform *open* operation with a vertically elongated structural element and compute connected components. From each connected component c, three human candidates can be generated assuming: 1) the centroid of the c is aligned with the center of human body; 2) the top center point of c is aligned with the human head; and 3) the bottom center point of c is aligned with the human feet.

3.2 Markov chain dynamics

Denote the state at iteration $t-1$ as $\theta_{t-1} = \{n, \{M_1, ..., M_n\}\}$. The following Markov chain dynamics are applied to θ_{t-1} which results in θ'. The dynamics correspond to sampling the proposal probability $q(\theta'|\theta_{t-1})$:

1. **Human hypothesis addition**: Randomly select a method from the three techniques described in Sec.3.1 to generate a new hypothesis. Assume that the hypothesized position is at (x_0, y_0), sample position (x_r, y_r) in a Gaussian density $N((x_0, y_0), diag(\sigma_x^2, \sigma_y^2))$. Rest of the parameters (l_r, h_r, f_r) are sampled from their respective prior distributions. A new human hypothesis is assembled as $M_{n+1} = \{l_r, x_r, y_r, h_r, f_r\}$. $\theta' = \{n+1, \{M_1, ..., M_n, M_{n+1}\}\}$.
2. **Human hypothesis removal**: Randomly select an existing human hypothesis $r \in [1, n]$ to remove. $\theta' = \{n-1, \{M_1, ...M_n\} - \{M_r\}\}$.
3. **Model/orientation switch**: switch the model/orientation label of a human hypothesis. Randomly select an existing human hypothesis $r \in [1, n]$ and randomly switch the model/orientation label l_r to another one. All other parameters are inherited.
4. **Stochastic diffusion of model parameters**: update the parameters of a human hypothesis in the direction of their gradients plus random noise. Randomly select an existing human hypothesis $r \in [1, n]$, update m_r according to:
   $$m_r \leftarrow m_r + k\frac{dE}{dm_r} + w$$
   where $E = logP(\theta|I)$, k is a step coefficient and w is a Gaussian noise. The noise helps avoid local maximums. The parameters are also bound to their minimum and maximum allowed values.
5. **Head position switch**: switch the head position of a human hypothesis. Randomly select an existing human hypothesis $r \in [1, n]$ and randomly switch the head position (x, y) to some other head candidate around the original position. It is equivalent to first removing this human and then adding the new one. However, removing r may result in a big decrease of the posterior probability so that it has small chance to be accepted. This is similar to building a bridge on a valley in the search space.

The first two are referred to as *jump* dynamics and the rest are referred to as *diffusion* dynamics. It is guaranteed that the Markov chain designed this way is *ergodic* (*i.e.*, any state is reachable from any other state within finite number of iterations) and *aperiodic* (*i.e.*, the Markov chain does not oscillate in a fixed pattern) since all of moves are stochastic [14]. Furthermore, redundant dynamics (*e.g.*, head position switch) is added for ease of traversal in the solution space. Multiple ways of adding human hypotheses increase the robustness of the system.

3.3 Incremental computation

In one iteration our algorithm only changes one object. Thus the new likelihood can be computed more efficiently by incrementally computing it only within the neighborhood of the area associated with this object and those overlapping with it. Furthermore, with the help of storing number of object layers at each pixel ($L(i)$), the incremental likelihood computation is only needed within the region of the object being changed.

When a human hypothesis is added, for each pixel i in the region of the added object,

```
if $L(i) = 0$ AND $F(i) = 0$
    $N_{10} \leftarrow N_{10} + 1;$
if $L(i) = 0$ AND $F(i) = 1$
    $N_{01} \leftarrow N_{01} - 1;$
$L(i) \leftarrow L(i) + 1;$
```

When a human hypothesis is removed, for each pixel i in the region of the removed object,

```
$L(i) \leftarrow L(i) - 1;$
if $L(i) = 0$ AND $F(i) = 0$
    $N_{10} \leftarrow N_{10} - 1;$
if $L(i) = 0$ AND $F(i) = 1$
    $N_{01} \leftarrow N_{01} + 1;$
```

In case of diffusion, N_{01} and N_{10} can be simply computed by a removal followed by an addition. The foreground residue map also gets updated incrementally at the same time.

Although a joint state and joint likelihood are used, the computation of each iteration is reduced to the region of an object (in a general case, to the region of an object's neighborhood) through the incremental computation. This is in contrast to particle filters where the evaluation of each particle (joint state) needs the computation of the full joint likelihood.

4 Experiments and Evaluations

We have tested the approach described above on a number of data sets and the results are very stable. Due to the space limit, we will mainly show the results and the performance evaluation on two sequences and mention the results of some other scenarios [1].

We perform background subtraction using a standard technique ([15]) and feed the foreground into the segmentation algorithm after morphology *close* operation. In our experiments, the parameters are fixed as the following: $\lambda_1 = 0.01$,$\lambda_2 = 0.005$,$\lambda_{10} = 0.8$, and $\lambda_{01} = 1.0$. Besides, we apply a hard constraint (25 pixels) on the minimum image height of human object. The Markov chain starts from a null state $\theta_0 = \{0, \emptyset\}$ and we assume $q(\theta_{t-1}|\theta') = q(\theta'|\theta_{t-1})$. The results are shown in input and output pairs. The outputs are the human models overlaid on the original images. Due to the small image size, we have explicitly marked the errors on the images when they occur: false alarms in black arrows and missed detections in white arrows.

Seq.1 is a 605-frame sequence captured from a camera on second floor with the camera tilt angle around 20^o. A group of 22 humans walked through the scene. The image sizes of human objects in the scene have a large variation (more than 10 times in area). The dense edges of the tree branch shadows at the far side result in high false positives for head candidates. There are 50 to 100 head candidates per frame. Result on the example frame in Fig.1 is given in Fig.5.

We compare our proposed approach with a random (uniform over the image) proposal probability for adding new human hypotheses in 5000-interation runs on the same frame. The log posterior probability histories of the two cases are shown in Fig.5.(b). The log posterior probability of our approach climbs quickly and then stays close to maximum value for fine adjustments while the log posterior probability of the random proposal probability improves significantly slower.

Our proposed approach is not sensitive to the initial state. We show the log posterior probability histories of 1000-iteration runs from a null initial state and an initial state containing 20 random humans on the foreground in Fig.5.(c). Although showing some difference at the beginning, they show little difference after 400 iterations.

Results on some more frames of Seq.1 are shown in Fig.7. The results were obtained by 2000-iteration runs.

Seq.2 is a 900-frame sequence captured from a camera above a building gate with the camera tilt angle $= 40^o$. A large tilt angle results in significant perspective effect on human shape in images. 33 humans passed by the scene with 23 going out of and 10 going in the building. Results on a few frames are shown in Fig.8. They were obtained by 1000-iteration runs.

To evaluate the accuracy of the method, we compare the result of each frame with ground truth derived from hand annotation. If a human object in the solution has an over 50% overlap with a human object in the ground truth, a match is declared. One-to-one mapping of the objects is enforced. The unmatched objects in the solution are declared as false alarms and the unmatched objects in the ground truth are declared as miss-detections. The humans whose bodies are partially outside the scene are not counted as mathches or errors (*i.e.*, they are “don’t cares”.). The results of the evaluation are summarized in Tab.1.

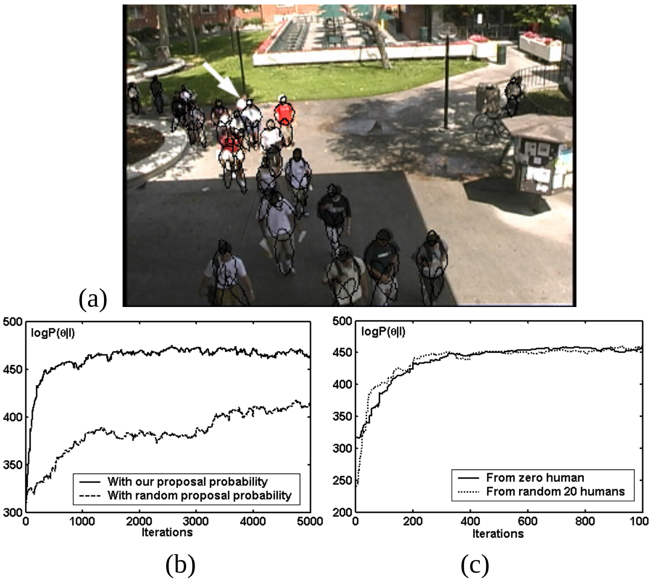
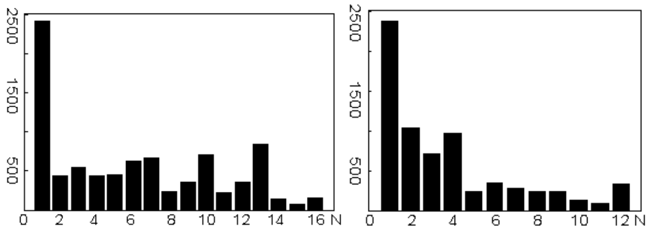


Figure 5: Experiment result of the image in Fig.1. (a) The result. False alarms are marked with dark arrows and miss detections are marked with white arrows (same for the following figures); (b) Compare convergence ($logP(\theta|I)$) with uniform proposal probability for adding human hypotheses; (c) Compare convergence from different initial states. Read text for detail.

To make the evaluation results more meaningful, we characterize the complexity of the dataset for human segmentation by the number of human objects per blob. Usually more humans per blob implies greater challenge to the segmentation algorithm. Fig.6 shows the histogram of the occurrence that a human object is in a blob of n ($n \in [1, \text{maximum number of objects in one blob}]$) human objects. In Seq.1, 50% of the humans appear in a blob containing 5 or more humans and 30% of the humans appear in a blob containing 9 or more humans. In Seq.2, about 30% of the humans appear in a blob containing 5 or more humans. The two sequences have similar detection rate. Seq.1 has a higher false alarm rate due to the foreground introduced by the morphology *close* operation when the humans are small (at the far side) and close to each other.

Besides the errors marked on the output, we show some other errors in Fig.9. Most of the missed detections are due to the failure of the background subtraction algorithm when the

[1] The results of all frames are available in video format at http://iris.usc.edu/˜taozhao/papers/CVPR03/CVPR03.html.



(a) (b)

Figure 6: The histogram of the number of human objects per blob in Seq.1 (a) and Seq.2 (b).

Table 1: Results of performance evaluations on Seq.1 and Seq.2.

	Seq.1	Seq.2
valid humans	8466	6726
correct detections	7881	6243
missed-detections	585	483
false alarms	291	12
detection rate	93.09%	92.82%
false alarm rate	3.43%	0.18%

human's clothing has color similar to the background or the overlapping humans form ambiguous shape or a combination of the two. The false alarms are mainly due to the inclusion of large (relative to a human's size) non-human regions in the foreground. These errors can be reduced by utilizing other image cues or temporal cues (*e.g.* tracking). Some of the selected model/orientations are not accurate. They are mainly due to the error of the foreground around the feet.

From the experiments we performed (not all shown here), we found that our proposed method is insensitive to blob fragmentation in cases with a significant amount of foreground indicating the presence of a human/humans. It also works robustly when there are significant amount of noises in the foreground.

The computation is affected by the complexity of the scene. More iterations are needed for a scene containing more humans and more occlusion. As an example, a 1000-iteration run on the above reported dataset requires about 0.5 seconds of CPU time on a Pentium IV 2.7G Hz PC, with un-optimized C++ code. If there are a small number of people in the scene, the system runs in real-time.

5 Conclusion and Future Work

We have presented an approach to segmenting individual humans in a crowded scene acquired from a static camera. The problem is formulated as a Bayesian MAP estimation problem and the solution is pursued using an efficient Markov chain Monte Carlo approach. The quality of the results is dependent on the definition of posterior probability according to image formation. Efficiency is obtained by incorporating domain knowledge as the proposal probability of the Markov chain. Experiments and evaluations on challenging real-life data show promising results. We feel that due to the Bayesian formulation, the described approach is more robust and effective in a wider range of scenarios (*e.g.* crowded situations as reported here) compared to earlier work (*e.g.* [16]).

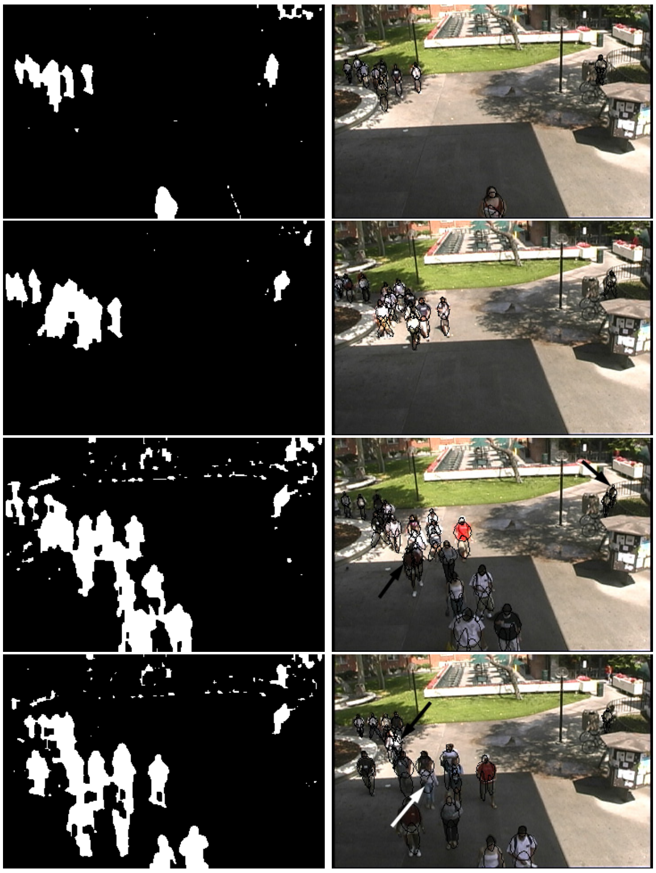


Figure 7: More results on the Seq.1. Left column: input; right column: output.

The work described here could be improved or extended in several ways. Currently the likelihood is based only on a binary foreground mask, other image cues such as edge or color could be used to reduce some ambiguities but at an increase in computation cost. Non-human objects (*e.g.* cars) should be added to the model set for more versatility but this too will increase computation and possibly result in more misclassifications. Finally, the work should be extended to include tracking: tracking information will provide important temporal priors which will both resolve some ambiguities of single frame analysis and reduce the computation. Tracking is also needed as an input for event inference algorithms.

Acknowledgement

The authors would like to thank Dr. Song-Chun Zhu and Dr. Zhuowen Tu for helpful discussion and the anonymous reviewers for their helpful comments.

References

[1] G. Borgefors, Distance Transformations in Digital Images, *CVGIP*, 34, 344-371, 1986.

[2] J. Canny, A Computational Approach to Edge Detection, *IEEE Trans. on PAMI*, Vol. 8, No. 6, Nov. 1986.

[3] A. M. Elgammal and L. S. Davis, Probabilistic Framework for Segmenting People under Occlusion, *Proc. Int. Conf. on Computer Vision*, Vancouver, Canada, 2001.

[4] D. M. Gavrila and V. Philomin, Real-time Object Detection for "Smart" Vehicles, *Proc. Int. Conf. on Computer Vision*, Kerkyra, Greece, 1999.

[5] S. Haritaoglu, D. Harwood and L. S. Davis, W4: Real-Time Surveillance of People and Their Activities, *IEEE Trans. on PAMI, Vol. 22, No. 8, 2000.*

[6] R. Hartley and A. Zisserman, *Multi View Geometry*, pp.191, Cambridge Press, 2000.

[7] F. Han, Z. Tu, and S. C. Zhu, A Stochastic Algorithm for 3D Scene Segmentation and Reconstruction, *Proc. European Conf. on Computer Vision; LNCS 2352*, 2002.

[8] M. Isard and J. MacCormick, BraMBLe: A Bayesian Multiple-Blob Tracker, *Proc. Int. Conf. on Computer Vision*, Vancouver, Canada, 2001.

[9] A. Mohan, C. Papageorgiou, and T. Poggio, Example-based Object Detection in Image by Components, *IEEE Trans. on PAMI*, vol.23, no. 4, Apr., 2001.

[10] A. Mittal and L. S. Davis, M2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene Using Region-Based Stereo, *Proc. of European Conf. on Computer Vision, LNCS 2350*, pp. 18-33, 2002.

[11] N. T. Siebel and S. Maybank, Fusion of Multiple Tracking Algorithm for Robust People Tracking, *Proc. European Conf. on Computer Vision; LNCS2353*, pp. 373-387, 2002.

[12] H. Tao, H. S. Sawhney and R. Kumar, A sampling algorithm for tracking multiple objects, *Proc. Workshop of Vision Algorithms, with ICCV 99*.

[13] L. Tierney, Markov chain concepts related to sampling algorithms, *Markov Chain Monte Carlo in Practice*, Chapman and Hall, pp. 59-74, 1996.

[14] Z. W. Tu and S. C. Zhu, Image Segmentation by Data-Driven Markov Chain Monte Carlo, *IEEE Trans. on PAMI*, vol.24, no.5, 2002.

[15] C.R. Wren, A. Azarbayejani, T. Darrell and A.P. Pentland, Pfinder: Real-time Tracking of the Human Body, *IEEE Trans. on PAMI*, Vol. 19, No. 7, 1997.

[16] T. Zhao, R. Nevatia and F. Lv, Segmentation and Tracking of Multiple Humans in Complex Situations, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Kauai, Hawaii, 2001.

[17] T. Zhao and R. Nevatia, Stochastic Human Segmentation from a Static Camera, *Proc. IEEE Workshop on Motion and Video Computing*, Orlando, Florida, 2002.

[18] S. C. Zhu, R. Zhang, and Z. W. Tu. Integrating Top-down/Bottom-up for Object Recognition by Data Driven Markov Chain Monte Carlo, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Hilton Island, North Carolina, 2000.

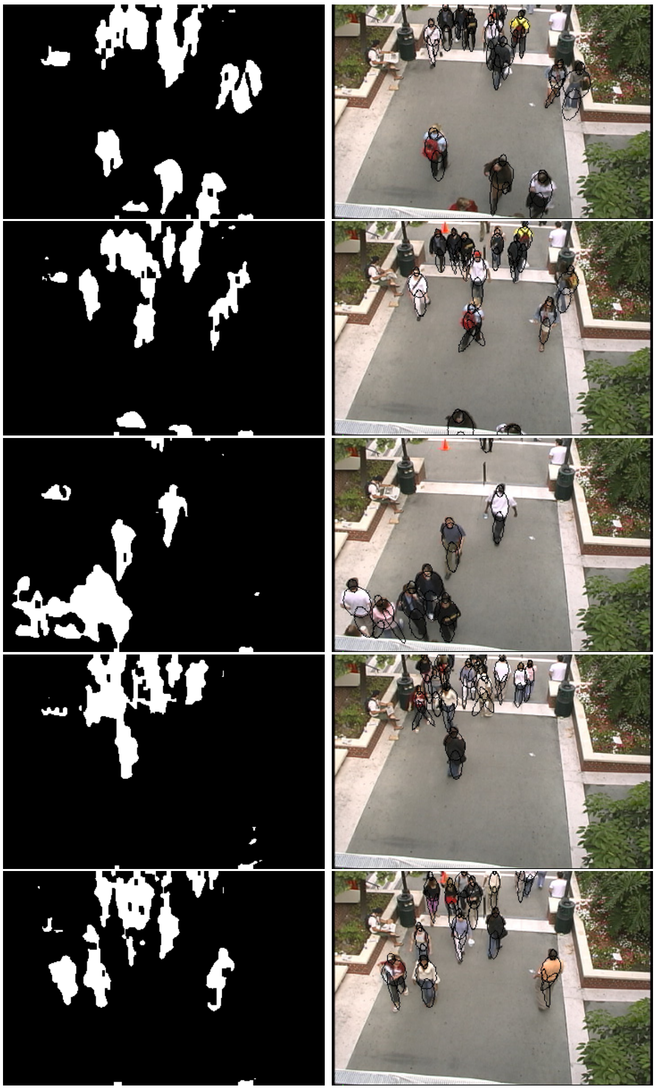


Figure 8: Selected frames of Seq.2. Left column: input; right column: output.

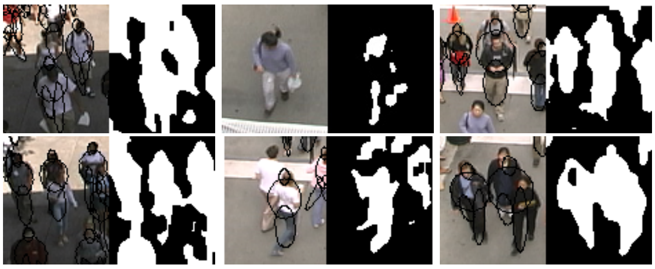


Figure 9: Some errors produced by the system.