Stochastic Human Segmentation from a Static Camera *

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

Abstract

Segmenting individual humans in a high-density scene (e.g., a crowd) acquired from a static camera is challenging mainly due to object inter-occlusion (Fig.1). We define this problem as a "model-based segmentation" problem and the solution is obtained using a Markov chain Monte Carlo (MCMC) approach. Knowledge of various aspects including human shape, human height, camera model, and image cues including human head candidates, foreground/background separation are integrated in a Bayesian framework. We show promising results on some challenging data.

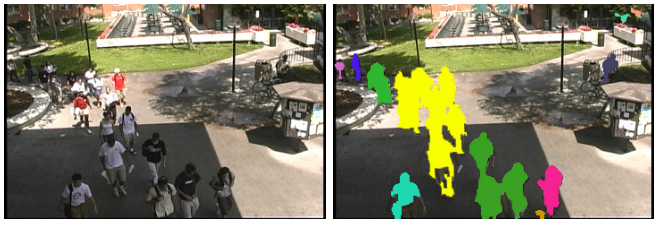
1 Introduction and Motivation

Segmentation and tracking of humans in video sequences is important for a number of tasks such as video surveillance and event inference as humans are the principal actors in daily activities of interest. We consider scenarios where the camera is fixed; in this case, moving pixels (foreground) can be detected fairly reliably by comparing the incoming frames with a background model.

However, the problem of obtaining object level descriptions from moving pixels begins to gain more attention. The difficulty of this problem increases with the density of the objects in the scene. We would like to categorize the characteristics of the problem according to the object density as follows:

- *Low density:* The moving objects are sparse and seldom overlapped. In this case, connected components of the foreground (called *blob*) correspond well to individual objects.
- *Moderate density:* Moving objects have simple interactions such as passing-by or small groups of people. In this case, using the knowledge of human head position and unexplained foreground ([11] [7]) or vertical projection of the blob ([4]) is able to solve the segmentation problem satisfactorily in most cases. Besides, a human specific model initialized before occlusion can also help solve the problem ([3] [11]).

[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.



(a) (b)

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

- *High density:* Blobs consist of large groups of people. People in the scene occlude each other either due to their spatial proximity or due to camera projection (an example frame and its foreground are shown in Fig.1). In this case, many of the heads are not on the foreground boundary and the vertical projection of the big blob is also not informative enough to perform segmentation.

Previous research on segmenting and tracking of multiple humans ([8] [4] [6] [3] [11] [7], etc) has been focused mainly on the first two classes. The high density case is interesting because such scenes usually contain rich human behaviors of interests. However the challenge is also obvious. Color segmentation is not likely to segment the individual humans; motion segmentation also won't give satisfactory result due to the similarity of the motion of individuals in a group and the non-rigid motion of human. Face detection may not be effective due to the arbitrary viewpoint (consider viewed from back) and the low image resolution.

1.1 Problem definition

Our goal is to segment overlapping humans from the foreground. We define the problem as a model-based segmentation problem. The difference from general image (region) segmentation is that we use object shape model to constrain the segmentation instead of enforcing some homogeneity (e.g., color, texture, etc) within the segments. The difference from object detection methods is that we know in advance that the image region of interest is made up of only the objects of known classes (or largely so) but the objects may occlude each other

(severely).

Assume we have an object model **M** and an image region I (foreground) is made up of an unknown number (N) of such objects with parameters $M_1, M_2, ..., M_N$, where M_i may contain position, size and other shape parameters. A solution θ is in the form of $\{N, \{M_1, ..., M_N\}\}$. Such a joint state has to be used due to their possible overlapping. Obviously, the solution space Σ is of unknown dimensions. The problem is to find a solution θ^* which maximize a posteriori probability (MAP):

$$\theta^* = argmax_{\theta \in \Sigma} P(\theta | I^{obs}) \quad (1)$$

where I^{obs} is various features that we can extract from the image. In other words, we need to find the number of humans and the location, size and other parameters for each human which maximize the posterior probability.

The posterior probability $P(\theta|I^{obs})$ is further decomposed into a likelihood term and a prior term according to Bayes rule:

$$P(\theta|I^{obs}) = \alpha P(I^{obs}|\theta)P(\theta) \quad (2)$$

where α is a normalization factor independent to θ.

1.2 Outline of our approach

Optimization in such a complex solution space is generally difficult. The Markov chain Monte Carlo (MCMC) provides a sampling method to traverse the complex solution space with a Markov chain and the optima can be found during the traversal. Recently, DDMCMC (data-driven MCMC) has been proposed in [12] and [10] for object recognition and image segmentation which emphasizes incorporating domain knowledge (heuristics) to design the proposal probability of the Markov chain to make the traversal more efficient compared to the traditional more random proposal probability. Our work is also motivated by this idea.

A block diagram of our system is shown in Fig.2. The foreground is computed by subtracting the background from the input image frames. The foreground is to be segmented into different overlapping human objects by the MCMC approach. We use a human model consisting of three parts to capture the gross shape of standing and walking humans. A prior based on the number of humans makes the segmentation concise (Sec.3.1). A joint likelihood based on the foreground/background separation is defined to minimize the difference of the real foreground and the foreground corresponding to the solution θ^* (Sec.3.4). Observing the head might be the only feature in the presence of severe inter-occlusion, we describe two techniques to detect heads based on the foreground boundary and intensity edges respectively (Sec.3.2). This is used as domain knowledge to direct the Markov chain on where to create new human hypotheses. Together with *removal* and *diffusion* processes, a well-balanced Markov chain dynamics is designed (Sec.3.3). We show promising experimental results on both outdoor and indoor sequences (Sec.4).

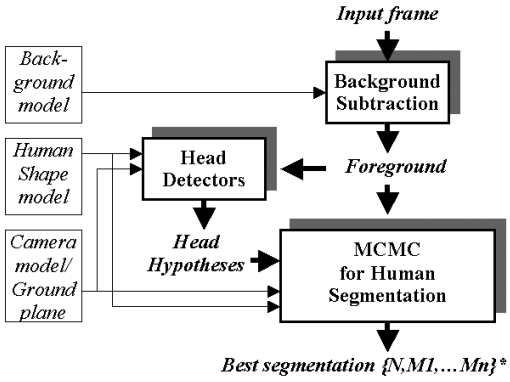


Figure 2: The block diagram of the system. Shaded box: program module; plane box: model; thick arrow: data flow; thin arrow: model association.

2 The MCMC Framework for MAP Estimation

Suppose we want to compute the maximum a posteriori value from a probability distribution $Q(\theta)$ (we use $Q(\theta) = P(\theta|I^{obs})$ for clarity), this is done by sampling from $Q(\theta)$ and finding the maximum value. To sample from $Q(\theta)$, a Markov chain can be designed. A widely used algorithm is the Metropolis-Hasting algorithm . At each iteration t, we sample a candidate state θ' from a *proposal distribution* $q(\theta_t|\theta_{t-1})$ (in simple words, what new state should the Markov chain go to from the current state.). The candidate state θ' is accepted with the following probability:

$$p = min(1, \frac{Q(\theta')q(\theta_t|\theta')}{Q(\theta_t)q(\theta'|\theta_t)})$$

If the candidate state θ' is accepted, $\theta_t = \theta'$, otherwise, $\theta_t = \theta_{t-1}$. It can be proven that the Markov chain constructed this way always has its stationary distribution equal to $Q()$, independent of the choice of the proposal probability $q()$ and the initial state θ_0 [9]. The advantage of the MCMC approach over other optimization methods is that with *jump* and *diffusion* dynamics it can work in complex solution space such as the one here with varying number of dimensions.

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 ([12] [10]) 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 widely used *hypothesis and testing* approach ([12]).

In our problem, head candidates which are detected from both the foreground boundaries and the intensity edges serve as domain knowledge on where to generate human hypotheses.

(a) (b)

Figure 3: 3-ellipsoid human model. (a) The model is composed of 3 parts: head, torso and legs; (b) The 2D heights computed by camera model from a fixed head-top assuming the 3D heights are 1.6, 1.7 and 1.8 meters respectively.

The components of the MCMC framework are described in detail in the next section.

3 Human Segmentation from Foreground Using MCMC

3.1 Human model and human count prior

Human body is highly articulated. To model it precisely, a complicated kinematics model with over 20 DOF is needed. However, in our application the human image is generally small and human motion is mostly limited to standing or walking. We can thus use a low dimensional model which is a composition of 3 fixed ellipsoids: one for the head, one for the torso and one for the legs (Fig.3.(a)). An ellipsoid fits these forms well and has a nice property that its projection is an ellipse with a convenient form [5]. The parameters of the human model include *position*, *height* and *fatness* ($M = \{x, y, h, f\}$). The height parameter also affects the width of the human shape proportionally and the fatness parameter captures the extra fatness besides the proportional change. We assume that human height has a uniform distribution in the range of $[1.5m, 1.9m]$ and the fatness is in the range of $[0.7, 1.3]$ times of the average one.

We also assume that the camera model and the ground plane are known. The camera model along with the height range can add constraints to the image height of a human. Fig.3.(b) shows the projected human model computed by the camera model with the 2D head position fixed assuming the 3D heights are 1.6, 1.7 and 1.8 meters respectively.

A prior on the number of humans in the scene is applied as $P(\theta)$ in Equ.2 to make the segmentation concise by penalizing redundant human hypotheses which do not contribute enough to the joint likelihood.

$$P(\theta) = P(N) = e^{-\lambda_1 N} \tag{3}$$

where λ_1 is a coefficient whose value is only meaningful relative to the coefficient in the likelihood measurement. Therefore, we set $\lambda_1 = 1$.

3.2 Head candidate detection

We observe that the head is almost the only reliably visible feature for a human in highly occluded environments observed from above. Therefore, we developed two techniques to detect head from image frames based on foreground boundary and the intensity edges respectively. The head candidates set $HC = HC_{boundary} \cup HC_{edge}$.

3.2.1 Head candidates from foreground boundary ($HC_{boundary}$)

This method detects the heads which are on the boundary of the foreground [11]. The basic idea is to find the local peaks of the boundary. The peaks are further filtered by checking if its corresponding human is higher than a minimum interested height and 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 blob. Fig.4.(a) shows the $HC_{boundary}$ on the example frame.

3.2.2 Head candidates from intensity edges (HC_{edge})

Here we describe a head detector based on image intensity edges which is also effective for the heads in the interior of the blob.

First Canny edge detection [2] is applied to the dilated foreground region of the input image. A distance transformation [1] is 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 the 3-ellipsoid 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. The detection score is also filtered in the same way as described in Sec.3.2.1. After smoothing it, we find all the peaks above a threshold which ensures an almost 100% detection rate but results in a high false alarm rate. An example is shown in Fig.4.(d). The

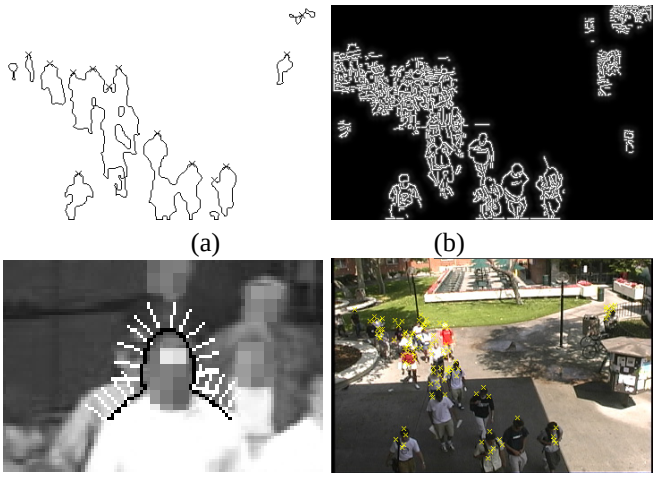


Figure 4: Head detectors. (a) Head detection from motion boundaries ($HC_{boundary}$); (b) Distance transformation on Canny edge detection result; (c) The head-shoulder (Ω) model: red-head part, blue-shoulder part, green-normals; (d) Head detection from intensity edges (HC_{edge}). See Sec.3.2 for detail.

false alarms tend to happen in the area of rich texture where there are abundant edges of various orientations.

3.3 Dynamics of the Markov chain

Denote the state at iteration $t-1$ as $\theta_{t-1} = \{N, \{M_1, ..., M_N\}\}$. The following Markov chain dynamics are designed which corresponds to sampling the proposal probability $q(\theta'|\theta_{t-1})$:

- Add a new human hypothesis. Randomly select a head candidate $hc \in HC$ which is not too close to any current heads. A new human hypothesis is assembled as $M_r = \{hc_x, hc_y, hc_h, hc_f\}$ where height hc_h and fatness hc_f are randomly assigned. $\theta' = \{N+1, \{M_1, ..., M_N, M_r\}\}$.
- Remove an existing human hypothesis. Randomly select an existing human r to remove. $\theta' = \{N-1, \{M_1, ...M_N\} - \{M_r\}\}$.
- Change the parameters of a human hypothesis. Randomly select an existing human r and add a random noise to its parameters. $M'_r = M_r + \{d_x, d_y, d_h, d_f\}$, $\theta' = \{N, \{M_1, ...M'_r, ...M_N\}\}$.

The first two are referred to as *jump* dynamics and the last one is referred to as *diffusion* dynamics. In each iteration, one of these is chosen randomly according to predefined probabilities P_{add}, P_{remove} and P_{change} ($P_{add}+P_{remove}+P_{change} = 1$) respectively. The above dynamics guarantee the Markov chain designed this way is *irreducible* (any state is reachable from any other state within finite number of iterations) and *aperiodic* (the Markov chain does not oscillate in a fixed pattern) since all of them are stochastic.

Figure 5: The likelihood based on the number of wrongly classified pixels. False positives: in white; true negatives: in shade. NOTE: single-ellipsoid model instead of 3-ellipsoid model is used for illustration.

3.4 Multi-human joint likelihood

3.4.1 Image likelihood by foreground/background separation

Since multiple humans may occlude each other, the image likelihood cannot be decomposed into image likelihood associated with each individual human. Here we use a simple joint likelihood based on foreground and background separation.

$$P(I_s^{obs}|\theta) = e^{-\lambda_2 N_{error}} = e^{-\lambda_2 (N_{fp}+N_{tn})} \quad (4)$$

Basically, the likelihood reflects the number of wrongly classified pixels (Fig.5; Equ.4) according to the current state and the foreground/background of the image frame. The wrongly classified pixels (N_{error}) include false positives (N_{fp}, as marked in Fig.5 in white region) and true negatives (N_{tn}, as marked in Fig.5 in shaded region). False positives refer to pixels that do not belong to the foreground but are within an hypothesized object; true negatives refer to pixels that belong to foreground but are not within any hypothesized object. One important feature of this likelihood is that it is invariant to depth ordering.

By combining Equ.3 and Equ.4, Equ.2 becomes:

$$P(\theta|I^{obs}) = \alpha e^{-\lambda_1 N} e^{-\lambda_2 (N_{fp}+N_{tn})} \quad (5)$$

λ_2 is a coefficient which controls the number of correctly classified pixels a human hypothesis has to contribute (excluding those contributed by others) to increase the posterior probability. It is mainly decided by the size of human that the application is interested in.

3.4.2 Incremental computation of joint likelihood

In one MCMC iteration only one object may change, so the computation of the likelihood can be carried out more efficiently by incrementally computing it only within the area associated with this object and those overlapping with it.

We show here how this incremental computation is done when creating a new human hypothesis; situation of removing a human or changing the parameter of a human is similar. Following the notation in Sec.3.3, Ψ_r is the neighborhood set of a new human r. It is obvious that $N'_{fp} \geq N_{fp}$ and $N'_{tn} \leq N_{tn}$.

$$P(I_s^{obs}|\theta') = e^{-\lambda_2 N'_{error}} = e^{-\lambda_2(N'_{fp}+N'_{tn})}$$
$$= e^{-\lambda_2(N_{fp}+\delta N_{fp}+N_{tn}-\delta N_{tn})}$$
$$= P(I_s^{obs}|\theta_{t-1}) * e^{-\lambda_2(\delta N_{fp}-\delta N_{tn})}$$

where δN_{fp} is the number of pixels which satisfy: 1) it is in human r; 2) it belongs to foreground; 3) it is not in any human $\in \Psi_r$. δN_{tn} is the number of pixels which satisfy: 1) it is in human r; 2) it belongs to background; 3) it is not in any human $\in \Psi_r$.

4 Experiment Results

We have tested the approach described above on several data sets and show the results on one outdoor sequence (as the example shown in Fig.1) and one indoor sequence. We set $P_{add} = 0.2$, $P_{remove} = 0.2$, and $P_{change} = 0.6$. In both cases, 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 pairs of inputs and outputs. The inputs are the foreground blobs and the head candidates; the outputs are the humans overlaid on the original images with semi-transparent randomly colored masks. Due to the small image size, we have explicitly marked the errors: false alarms in black arrows and miss detections in white arrows.

The outdoor sequence is captured from a camera on second floor. At the far end the humans are very small and the dense edges result in high false alarm rate of head candidates. There are 50 to 100 head candidates per frame. We set $\lambda_2 = 0.02$ and 5000 iterations are used. Result on the example frame in Fig.1 is given in Fig.6. Fig.6.(c) shows the log posterior probability of all the iterations. We can see that the posterior climbs very fast and then stays close to optimal solution for fine adjustment. Fig.7 shows the results for more frames in the same sequence. The errors usually happen at the far end and some of them are also difficult for a human observer.

The indoor sequence has a lower viewpoint (camera mounted on a tripod on the ground) and the inter-occlusion is more severe. The average number of head candidates is around 100. We set $\lambda_2 = 0.004$ and 5000 iterations are used. Fig.8 shows the results on some of the frames. The people in the scene cause significant illumination changes both on the carpet and on the pillars which causes our background subtraction method to give erroneous foreground blobs. We manually removed those false foreground detections to isolate the difficulty of human segmentation. However, the head candidates were obtained from the real foreground. The missed detections either have too few pixels or are almost completely overlapped with other humans.

The computation is affected by the complexity of the scene. More iterations are needed for a scene containing more humans and more occlusion. The computation required for each iteration is also proportional to the size of the foreground region and the number of human objects in the state. As an example, a 5000-iteration run on the above reported dataset

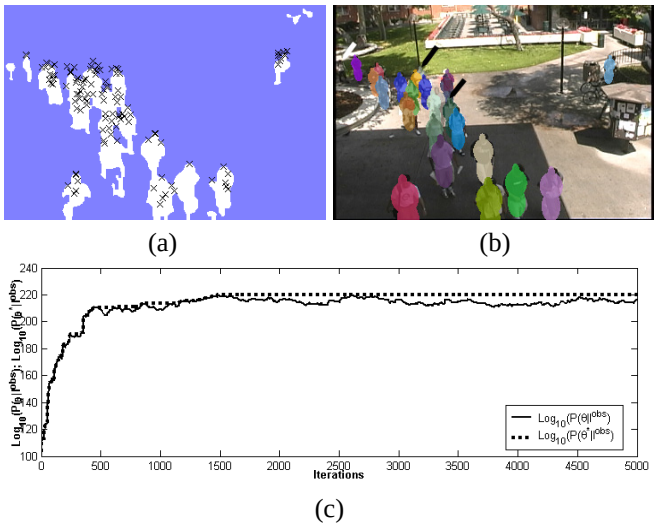


Figure 6: Experiment result of the image in Fig.1. (a) The input; (b) The result. False alarms are marked with dark arrows and miss detections are marked with white arrows (same for the following figures); (c) $log_{10}P(\theta|I^{obs})$ (solid line) and $Log_{10}P(\theta^*|I^{obs})$ (dotted line) over iterations.

requires about 15 seconds on a Pentium 4 1.5G Hz PC, with un-optimized C++ code. However, we have observed in both sequences, 1000 iterations are sufficient for the human objects of larger size to be segmented correctly (as can be seen in Fig.6.(c)).

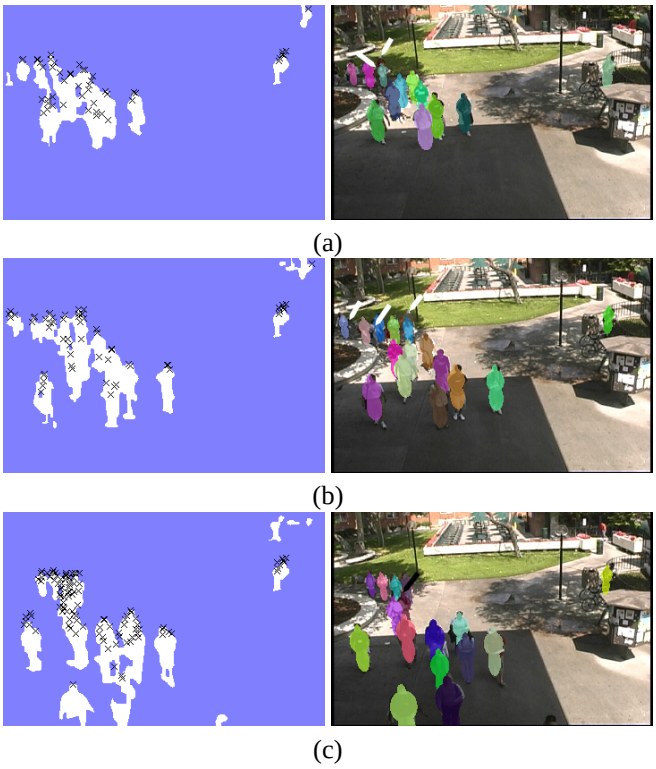
5 Conclusion and Future Work

We have presented an approach to segment individual humans in a high-density scene (e.g., a crowed) acquired from a static camera. This problem is made difficult by objects being inter-occluded. We define the problem as a *model-based segmentation* problem where a human shape model is used. A posterior probability is defined based on a human count prior and a foreground/background separation likelihood function. The MAP solution is found using an MCMC framework. Human head candidates are computed from both the foreground boundaries and the intensity image edges to direct the human hypothesizing process of the Markov chain dynamics. Experimental results have shown that this approach is highly promising.

In this work, knowledge of various aspects including human shape, human height, camera model, and image cues including human head candidates, foreground/background separation are integrated in a Bayesian framework. Same method can also be used in other similar problems.

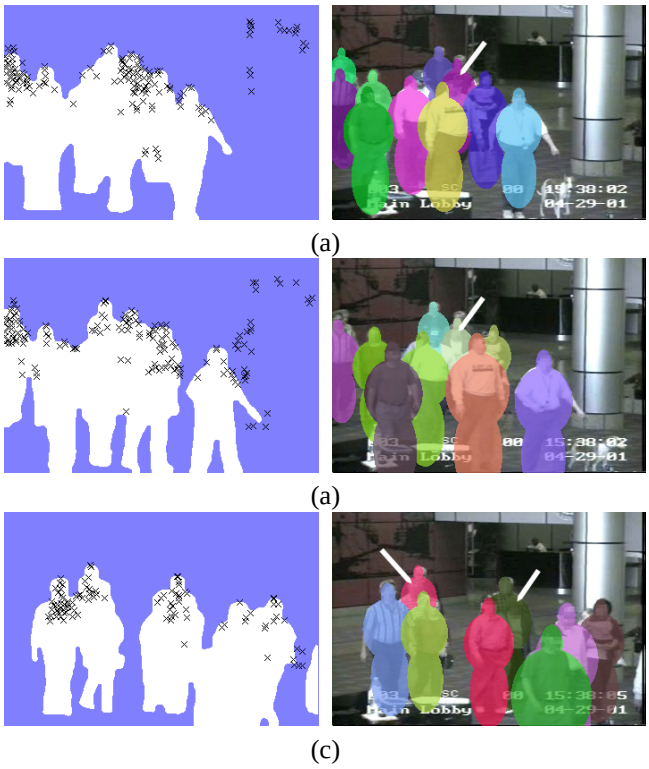
5.1 Future work

The work described here can be improved in various aspects easily since the MCMC is an open framework. We just list a few important ones here:



(a) (b) (c)

Figure 7: More results on the outdoor sequence. Left column: input; right column: output.



(a) (a) (c)

Figure 8: Results on the indoor sequence. Left column: input; right column: output.

- Some widely used techniques can be implemented in the Markov chain dynamics such as having more jump steps in the early stage and having more diffusion steps in the later stage; deciding which object to remove according to a fitness value, etc.
- Currently we only use foreground/background separation as likelihood measurement. It will not be effective if many people are entirely in the interior of the foreground. Other likelihood cues (e.g. edge, color and motion) may be helpful.
- Currently we don't use any temporal information except for background subtraction. The framework should be extended to make temporal estimation.
- Human model can be enhanced to handle closer range cases where the articulation of the limbs cannot be well captured by our current model. However, the low dimensionality of the model has to be ensured.

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. of Int. Conf. on Computer Vision*, Vancouver, Canada, 2001.

[4] 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.*

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

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

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

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

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

[10] 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.

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

[12] 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. of IEEE Conf. on Computer Vision and Pattern Recognition*, Hilton Island, North Carolina, 2000.