Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

# Tracking multiple people under occlusion and across cameras using probabilistic models

Xuan-he WANG[1,2†], Ji-lin LIU[1,2]

([1]*Institute of Information and Communication Engineering, Zhejiang University, Hangzhou 310027, China*)

([2]*Zhejiang Provincial Key Laboratory of Information Network Technology, Hangzhou 310027, China*)

[†]E-mail: xuanhewang@163.com

Received June 21, 2008; Revision accepted Oct. 24, 2008; Crosschecked Apr. 27, 2009

**Abstract:** Tracking multiple people under occlusion and across cameras is a challenging question for discussion. Furthermore, the cameras in this study are used to extend the field of view, which are distinguished from the same field of view. Such correspondence between multiple cameras is a burgeoning research subject in the area of computer vision. This paper effectively solves the problems of tracking multiple people who pass from one camera to another and segmenting people under occlusion using probabilistic models. The probabilistic models are composed of blob model, motion model and color model, which make the most of the space, motion and color information. First, we present a color model that uses maximum likelihood estimation based on non-parametric kernel density estimation. Second, we introduce a blob model based on mean shift, which segments the body into many regions according to the color of each person in order to spatially localize the color features corresponding to the way people are dressed. Clothes can be any mixture of colors. Third, we bring forward a motion model based on statistical probability which indicates the movement position of the same person between two successive frames in a single camera. Finally, we effectively unify the three models into a general probabilistic model and attain a maximization likelihood probability image, which is used to segment the foreground region under occlusion and to match people across multiple cameras.

**Key words:** Color model, Motion model, Blob model, People occlusion, People tracking, Kernel density estimation
**doi:**10.1631/jzus.A0820474 **Document code:** A **CLC number:** TP391

## INTRODUCTION

In realistic visual surveillance scenarios, an automated vision system aimed at consistently tracking humans has been a challenging topic in computer vision. However, the most complex challenge is to deal with the occlusions of people in a group and to track the people across cameras with an overlapping field of view (Khan and Shah, 2003).

In related work, two approaches are mainly adopted during the course of tracking: monocular approaches and multi-view approaches. Monocular approaches include the color-based monocular methods (Mittal and Davis, 2003; Kang *et al*., 2004) and the blob-based monocular methods (Black *et al*., 2002). Multi-view approaches include the color-based multi-view methods (Khan and Shah, 2000; Smith *et al*., 2005) and the blob-based multi-view methods (Han *et al*., 2004). Tracking people in a single view is prior to tracking in multi-views across cameras. The Bramble system (Isard and MacCormick, 2001) is a multi-blob tracker that generates a blob-likelihood which is computed using a variation on the Bayesian correlation scheme presented by Sullivan *et al*.(1999). Then it achieves the people tracking by means of particle filter on condition that the number of objects is unknown. Romano *et al*.(2000) presented an approach for tracking in cameras with an overlapping field of view that did not require calibration. The camera calibration information was recovered by matching motion trajectories obtained from different views, and plane cosmographies were computed from the most frequent matches. The approach of (Khan and Shah, 2006) was to use planar homographs constraint to resolve occlusions and to determine locations on the ground plane

corresponding to the feet of the people in crowds by means of many cameras in the same field of view. Du and Piater (2007) presented an approach to tracking people in multiple cameras and to making use of both each camera and ground plane by individual particle filters. Firstly, the particle filters in each camera pass messages to those in the ground plane where the multi-camera information is integrated by intersecting the target's principal axes to relax the dependence on precise foot positions when mapping targets from images to the ground plane using homography. Secondly, the fusion results in the ground plane are then incorporated by each camera as boosted proposal functions. In (Otsuka and Mukawa, 2004), a recursive Bayesian estimation approach was used to deal with occlusions while tracking multiple people in multi-view. The algorithm tracks objects located in the intersections of 2D visual angles, which are extracted from silhouettes obtained from different fixed views. When occlusion ambiguities occur, the recursive Bayesian estimation is applied, which includes two processes: one is hypothesis generation according to predicted object states and previous hypotheses and the other is hypothesis testing using a branch-and-merge strategy.

## OUR APPROACH

This study presents a probabilistic 2D tracking of multiple people under occlusion and across cameras by taking advantage of the available probabilistic model including blob model, motion model and color model. The cameras in this study are used to extend the field of view, which are distinguished from the same field of view. People are being tracked by different cameras by a maximum likelihood method based on an appearance model. The appearance model includes a blob model based on mean shift and a color model based on Gaussian kernel density. The motion model is to further help verify the people being tracked with one camera. The motion model based on Gaussian probability indicates the movement position of the same person between two successive frames in a single camera. We effectively unify the blob model, color model and motion model into a general probabilistic model to track people. The occlusion is also handled by means of the maximum likelihood method. Our system does not require any inter-camera calibration. The probabilistic models of this study are generic and applicable to many situations. The probabilistic models are as follows:

(1) A background model builds a statistical model for a background scene, which is used to detect foreground regions. Then we focus on the detection of people from foreground regions using shape and area cues.

(2) A blob model segments the body into many regions according to the color of each person in order to spatially localize the color features corresponding to the way people are dressed. The goal of the blob model is to cluster the similar colors together.

(3) A motion model is used to estimate the movement position of the same person in successive frames in a single camera using statistical probability. The same person in successive frames has the nature of a larger probability value in the course of the movement of people.

(4) A color model involves modeling color distribution and space distribution with respect to the body. We use a non-parametric approach based on kernel density estimation to estimate the color distribution of each blob. Therefore, we do not restrict the clothing to be of uniform color. It can be any mixture of color.

The overview block diagram in Fig.1 shows the system architecture of the proposed method. In the first stage, we reconstruct the background of the video with the Gaussian mixture background model (Stauffer and Grimson, 1999). Then we detect foreground pixels using background subtraction and remove the shadow by means of the algorithm in (Cucchiara *et al*., 2003). In the second stage, if the foreground persons are of the first appearance, we will store them in stored model, through the block of a new person. If the foreground persons are not of the first appearance, we will store them in the stored model, through the block of an update model. To attain a stored model is the initialization stage for tracking people. In the third stage, which is the most important stage of all, we will solve the problem of tracking multiple people under occlusion or across cameras by means of a blob model, color model and motion model. A blob model segments the people in the stored model into many sub-blobs according to the color of each person. Each sub-blob will be used by

the color model. So the blob model is the premise of the color model. In other words, the blob model serves the color model. We do not use the homography among cameras. Instead, we make the most of the color information, space information and motion information. Our algorithm optimizes space information and color information and implements the cluster using the advantage of gradient optimization of the mean shift. We track the people by means of the maximization likelihood image, which is obtained by the color density function and motion density function on the basis of the cluster. Finally, the tracking of humans under occlusion and across cameras will be implemented by the maximization likelihood images. The correspondence results are fed back to the stored model through the update model as described in Fig.1.

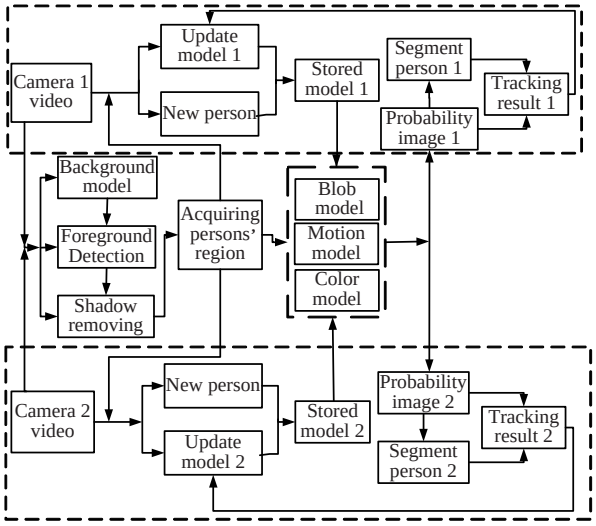



**Fig.1 System architecture of the proposed method**

## APPEARANCE MODEL

### Representation

The appearance model in this study is different from other appearance models (Lim *et al*., 2006; Yu *et al*., 2007). In this study, the appearance model consists of a blob model and a color model. In practice, people can be dressed in many different ways, but generally the way people are dressed leads to a set of major color regions. We decompose each person in the stored model into many blobs $M'=\{A_i\}$ according to the color of clothes. Therefore, we do not restrict the clothing to be of uniform color. In addition, we also pay attention to the space information. Supposing we attain a pixel from the black hair in the current person model, it will not match with the blob of the black shoes in the stored model. Modeling these persons involves modeling their color distributions and their spatial distribution with respect to the body. We consider here the case of both color and space.

First, we decompose a person into many blobs $M'=\{A_i\}$ using mean shift. We cluster the similar color of clothes based on mean shift. The count of blobs is determined by giving the parameters of kernel bandwidth and minimum pixel counts. If the count of small blobs is less than minimum pixel count, we will merge a small blob with a nearby big and color-similar blob. For instance, we suppose that a person is divided into six blobs in Fig.2.

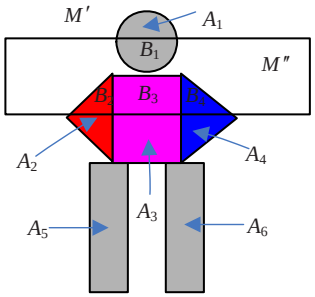



**Fig.2 Blob model according to color of clothes**

Then, we get a space set $M''$ relative to a pixel $\boldsymbol{c}_{t,x,y}$ position from the current person. The set can be expressed as

$$M''=\{(x,y)\,|\,0<x<w_{t-1},\ h_{t-1}(y_t/h_t-\alpha)<y<h_{t-1}(y_t/h_t+\alpha)\},$$

where $w_{t-1}$, $h_{t-1}$ represent the width and height of one of the persons of the last frame $t-1$, respectively, $h_t$ represents the height of a person who includes the pixel $\boldsymbol{c}_{t,x,y}$ in the current frame $t$, $y_t$ represents the $y$ axis coordinate of $\boldsymbol{c}_{t,x,y}$ in the current frame $t$, and $\alpha$ is a decimal fraction which means a variational range. The set $M''$ is changed as the current pixel $\boldsymbol{c}_{t,x,y}$ position is changed.

Then, we can get an intersection $M$ of $M'$ and $M''$,

$$M=\{B_1=A_1\cap M'',\ B_2=A_2\cap M'',\ B_3=A_3\cap M'',\ B_4=A_4\cap M'',\ldots\}.$$

The advantage of the intersection $M$ is to efficiently utilize the space information.

The blob model is finished, whose function aims at attaining the sub-sets $B_1$, $B_2$, $B_3$, $B_4$, …. Each blob $B_i$ will be used in the color model. Each blob $B_i$ has color density function $P_B(\boldsymbol{c}_{t,x,y})$. It can be seen from this that the blob model is the premise of the color model.

At last, given a set $M_{t-1,k}=\{B_1, B_2, B_3, B_4, \ldots\}$ correspondent to $\boldsymbol{c}_{t,x,y}$, the probability of the current pixel $\boldsymbol{c}_{t,x,y}$ which belongs to $M_{t-1,k}$ can be defined as Eq.(1), which is the appearance model:

$$\begin{aligned} &P(\boldsymbol{c}_{t,x,y} \mid M_{t-1,k}) \\ &\quad = \max\{P(\boldsymbol{c}_{t,x,y} \mid B_1), P(\boldsymbol{c}_{t,x,y} \mid B_2), P(\boldsymbol{c}_{t,x,y} \mid B_3), \ldots\} \\ &\quad = \max\{P_{B_1}(\boldsymbol{c}_{t,x,y}), P_{B_2}(\boldsymbol{c}_{t,x,y}), P_{B_3}(\boldsymbol{c}_{t,x,y}), \ldots\}. \end{aligned} \tag{1}$$

We consider the likelihood maximization in all the subsets $M_{t-1,k}=\{B_1, B_2, B_3, B_4, \ldots\}$ as the probability of $\boldsymbol{c}_{t,x,y}$. Each blob $B_i$ has the same color distribution which can be expressed as color density function $P_B(\boldsymbol{c}_{t,x,y})$. How to express the color density function of a person in virtue of $P_B(\boldsymbol{c}_{t,x,y})$ is the key technology in the whole course of tracking a person under occlusion or across cameras.

So far, we have obtained the probability of the appearance model $P(\boldsymbol{c}_{t,x,y}|M_{t-1,k})$ relative to a certain person $k$ in the last frame $t-1$. We will discuss in detail how to automatically attain the blob model in the next subsection 'Blob model' and how to obtain probability of the color density function in subsection 'Color model'.

**Blob model**

We decompose each person in the stored model into many blobs $M'=\{A_i\}$ according to the color of clothes using mean shift. The mean shift algorithm (Comaniciu and Meer, 1999) is a non-parametric clustering technique which does not require prior knowledge of the number of clusters, and does not constrain the shape of the clusters. We must make the most of the space and color information in practical applications. The space of the pixel in the image is considered as the spatial domain, while the color information is regarded as range domain. Thus, the space and color vectors are concatenated in the joint spatial-range domain $\boldsymbol{x}=(\boldsymbol{x}^{\mathrm{s}}, \boldsymbol{x}^{\mathrm{r}})$. The feature vector $\boldsymbol{x}=(\boldsymbol{x}^{\mathrm{s}}, \boldsymbol{x}^{\mathrm{r}})$ is not normalized before using mean shift iterations. The vectors $\boldsymbol{x}^{\mathrm{s}}$ and $\boldsymbol{x}^{\mathrm{r}}$ stand for the spatial part and range part of a feature vector, respectively. They have their own kernel bandwidth parameters $h_{\mathrm{s}}$ and $h_{\mathrm{r}}$, in the aspect of which we employed the fixed parameters. As long as the color-similar pixels are clustered, this can meet our requirement because our algorithm can adapt the change of blob count. We choose the space of $L^*u^*v^*$ as the range part of the color vector. The multivariate kernel is defined as the product of two radial symmetric kernels:

$$K_{h_{\mathrm{s}},h_{\mathrm{r}}}(\boldsymbol{x}) = \frac{C}{h_{\mathrm{s}}^2 h_{\mathrm{r}}^3} k\left(\left\|\frac{\boldsymbol{x}^{\mathrm{s}}}{h_{\mathrm{s}}}\right\|^2\right) k\left(\left\|\frac{\boldsymbol{x}^{\mathrm{r}}}{h_{\mathrm{r}}}\right\|^2\right), \tag{2}$$

where $k(x)$ obeys the Epanechnikov kernel which can provide satisfactory performance.

$$k(x) = \begin{cases} 1-x, & 0 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Its derivation is

$$g(x) = -k'(x) = \begin{cases} 1, & 0 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The deduction of the multivariate kernel has a similar deduction to the method of the univariate kernel (Comaniciu and Meer, 2002), so the mean shift vector of the multivariate kernel of spatial domain and range domain can be expressed as

$$\boldsymbol{y}_{j+1} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i g\left(\left\|\frac{\boldsymbol{y}_j^{\mathrm{s}} - \boldsymbol{x}_i^{\mathrm{s}}}{h_{\mathrm{s}}}\right\|^2\right) g\left(\left\|\frac{\boldsymbol{y}_j^{\mathrm{r}} - \boldsymbol{x}_i^{\mathrm{r}}}{h_{\mathrm{r}}}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\boldsymbol{y}_j^{\mathrm{s}} - \boldsymbol{x}_i^{\mathrm{s}}}{h_{\mathrm{s}}}\right\|^2\right) g\left(\left\|\frac{\boldsymbol{y}_j^{\mathrm{r}} - \boldsymbol{x}_i^{\mathrm{r}}}{h_{\mathrm{r}}}\right\|^2\right)}. \tag{5}$$

According to Eq.(5), the procedure of the blob model will be shown as follows:

**Algorithm 1** Mean shift

Input: $\boldsymbol{x}_i = (\boldsymbol{x}_i^{\mathrm{s}}, \boldsymbol{x}_i^{\mathrm{r}})$, $i=1, 2, \ldots, n$: Inputting pixels in the joint spatial-range domain (Fig.3a),

$h_{\mathrm{s}}$ and $h_{\mathrm{r}}$: Kernel bandwidth parameters of spatial domain and range domain, respectively,

$M$: Minimum pixel counts of set $A_p$. In Fig.3, the kernel bandwidth parameters and minimum pixel counts are ($h_s$, $h_r$, $M$)=(7, 11, 60).

Middle results: $\boldsymbol{y}_{i,c}=(\boldsymbol{y}_{i,c}^s, \boldsymbol{y}_{i,c}^r)$: The $i$th pixel in the joint spatial-range domain will converge to $\boldsymbol{y}_{i,c}$,

$\boldsymbol{z}_i=(\boldsymbol{z}_i^s, \boldsymbol{z}_i^r)$, $i$=1, 2, …, $n$: Storing convergence $\boldsymbol{y}_{i,c}$ of each pixel $\boldsymbol{x}_i$ into $\boldsymbol{z}_i$ (Fig.3b),

$\{C_p\}$, $p$=1, 2, …, $k$: Grouping together all $\boldsymbol{z}_i$ which are closer than $h_s$ and $h_r$ (Fig.3c).

Output: $M'=\{A_p\}$, $p$=1, 2, …, $k$: The set $M'$ of a stored person consisting of the blob subset $A_p$ (Fig.3d).

For $i$=1 to $n$

$\boldsymbol{y}_{i,1}=\boldsymbol{x}_i$;

For $j$=1 to $c$ (until convergence)

Compute

$$\boldsymbol{y}_{i,j+1}=\frac{\sum_{i=1}^{n}\boldsymbol{x}_i g\left(\left\|\frac{\boldsymbol{y}_{i,j}^s-\boldsymbol{x}_i^s}{h_s}\right\|^2\right) g\left(\left\|\frac{\boldsymbol{y}_{i,j}^r-\boldsymbol{x}_i^r}{h_r}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\boldsymbol{y}_{i,j}^s-\boldsymbol{x}_i^s}{h_s}\right\|^2\right) g\left(\left\|\frac{\boldsymbol{y}_{i,j}^r-\boldsymbol{x}_i^r}{h_r}\right\|^2\right)}.$$

$\boldsymbol{z}_i=\boldsymbol{y}_{i,c}$;

Get $k$ clusters $\{C_p\}$, $p$=1, 2, …, $k$ by grouping together all $\boldsymbol{z}_i$ which are closer than $h_s$ and $h_r$;

For $i$=1 to $n$

$L_i=\{p|\boldsymbol{z}_i\in C_p\}$;

Merge spatial regions containing less than $M$ pixels.

$A_p=\{\boldsymbol{x}_j|L_j=p, j=1, 2, ..., n\}$, $p$=1, 2, …, $k$;

So, $M'=\{A_p\}$.

(a) (b) (c) (d)

**Fig.3 Process of the blob model**

(a) Original image; (b) Filtered image; (c) Grouping the filtered image together; (d) Segmenting the original image into many blobs

The blob model algorithm is different from the mean shift algorithm for segmentation presented by Comaniciu in respect to the applied environment. The mean shift algorithm is applied to images. The blob model is applied to video. So we must utilize motion information of the video. Thus, we detect foreground pixels using motion information in advance. And then the foreground is segmented by means of the mean shift. In addition, the bandwidths $h_s$ and $h_r$ will affect the segmentation effect in the mean shift algorithm. However, our algorithm for tracking people is not sensitive to bandwidth $h_s$ or $h_r$, because as long as the color-similar pixels are clustered, this can meet our requirement. The count of blobs will not affect the tracking result greatly. At last, the algorithm of the mean shift is part of the blob model. Each blob of set $M'=\{A_i\}$ is attained by means of the mean shift. Then we will use the space information to attain the set $M''$ above mentioned. To attain the sub-sets $B_1$, $B_2$, $B_3$, $B_4$, …, which are attained by intersection $M$ of $M'$ and $M''$, is the intention of the blob model.

**Color model**

Color is one of the most important bits of information in tracking an object. We use the method of kernel density estimation (Giné *et al.*, 2004) to model the color density. Kernel density estimation belongs to a class of estimation called non-parametric density estimation. The advantage of non-parametric estimators is that they have no fixed structure and depend upon all the data points to reach an estimate. So we apply it to tracking people.

Let $B=\{x_i\}$, $i$=1, 2, …, $N$ be a random sample from a univariate distribution with unknown density $f$. Let $K$ be a symmetric probability density function and $\sigma$ be its scaling parameter or bandwidth. Then the standard kernel density estimation $f$ at a point $x$ is given by

$$\hat{f}(x)=\frac{1}{N\sigma}\sum_{i=1}^{N}K\left(\frac{x-x_i}{\sigma}\right). \quad (6)$$

If $x$ is changed into a $d$-dimensional vector, kernel density estimation can be achieved by the multivariate product kernel estimate (Scott and Sain, 2004). So Eq.(6) is changed into the following:

$$\hat{f}(\boldsymbol{x})=\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{\sigma_1\sigma_2\ldots\sigma_d}\prod_{k=1}^{d}K\left(\frac{\boldsymbol{x}^{(k)}-\boldsymbol{x}_i^{(k)}}{\sigma_k}\right)\right]. \quad (7)$$

Each dimension has a different smoothing parameter $\sigma_k$ in Eq.(7). We describe a color model as a function of multivariate product kernel estimation. In

order to handle illumination changes, we represent the normalized color of each pixel as a 3D vector $\boldsymbol{x}=\{r, g, s\}$, where $r=R$, $g=G$, $s=B$ or $r=R/(R+G+B)$, $g=G/(R+G+B)$, $s=(R+G+B)/3$, i.e., $d=3$ in Eq.(7). Given a sample set $B=\{\boldsymbol{x}_i\}$, $i=1, 2, \ldots, N$, which we get from the result of the blob model, the color density function of the current pixel $\boldsymbol{x}=\{r, g, s\}$ can be calculated as Eq.(8) based on kernel density estimation.

$$P_B(r,g,s) = \frac{1}{N\sigma_r\sigma_g\sigma_s}\sum_{i=1}^{N} K\left(\frac{r-r_i}{\sigma_r}\right)K\left(\frac{g-g_i}{\sigma_g}\right)K\left(\frac{s-s_i}{\sigma_s}\right), \quad (r_i,g_i,s_i)\in B. \tag{8}$$

In our method, a Gaussian kernel is used, i.e.,

$$K(s)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{s}{\sigma}\right)^2\right]. \tag{9}$$

Kernel density estimation is the canonical statistical method for the fundamental problem of estimating general probability densities without distribution assumptions.

More formally, kernel estimators smooth out the contribution of each observed sample $B=\{\boldsymbol{x}_i\}$, $i=1, 2, \ldots, N$ over a local neighborhood of that data point. The contribution of observed sample $B=\{\boldsymbol{x}_i\}$, $i=1, 2, \ldots, N$ to the estimate at some point $\boldsymbol{x}=\{r, g, s\}$ depends on how far apart each sample $B=\{\boldsymbol{x}_i\}$, $i=1, 2, ..., N$ and $\boldsymbol{x}=\{r, g, s\}$ are. The extent of this contribution is dependent upon the shape of the kernel function adopted and the bandwidth. More general bandwidth choices for multivariate density estimation are described in (Duong and Hazelton, 2005).

## MOTION MODEL

The motion model in this study is different from other motion models (Vega and Sarkar, 2003; Angel *et al.*, 2005). The motion model in this research is based on statistical probability which indicates the movement position of the same person between two successive frames in a single camera. The purpose of the motion model is to advance the exactness of tracking people. In practice, some persons wear similar clothes, so if the motion information can be used, this problem can be solved. Median coordinate vector $\boldsymbol{v}_t$ is the foundation of the motion model. Firstly, we attain a foreground image of one person using background subtraction. Then, we get a circum-rectangle and barycenter $(x_2, y_2)$ of the foreground image. Lastly, we do a perpendicular line and get two points of intersection: top point $(x_1, y_1)$ and bottom point $(x_3, y_3)$ (Fig.4). So the median ordinate vector $(x_2, y_2)$ can be expressed as the vector $\boldsymbol{v}_t=(x_1, y_1, x_2, y_2, x_3, y_3)$. The motion model is implemented by the motion density function $P_{M_{t-1}}(\boldsymbol{v}_t)$, which makes the most of space probability related to Euclidean distance. We use the median coordinate of each foreground region as an estimate of the object position in the image coordinate system (Fig.5). Fig.5 shows the corresponding median coordinate vector of the same person in succession.

The motion density function $P_{M_{t-1}}(\boldsymbol{v}_t)$ can be estimated by a joint Gaussian probability distribution function. These three elements in vector $\boldsymbol{v}_t$ is independent one another. So the motion density function $P_{M_{t-1}}(\boldsymbol{v}_t)$ can be expressed as

$$\begin{aligned} P_{M_{t-1}}(\boldsymbol{v}_t) &= g_{\sigma_1}(x_{t,1}-x_{t-1,1}, y_{t,1}-y_{t-1,1}) \\ &\times g_{\sigma_2}(x_{t,2}-x_{t-1,2}, y_{t,2}-y_{t-1,2}) \\ &\times g_{\sigma_3}(x_{t,3}-x_{t-1,3}, y_{t,3}-y_{t-1,3}), \end{aligned} \tag{10}$$

where

$$g_\sigma(x,y)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x^2+y^2}{\sigma^2}\right)\right]. \tag{11}$$

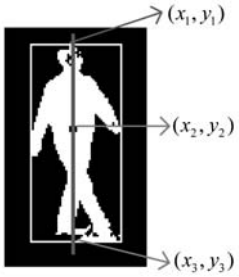



**Fig.4 Median coordinates of foreground**

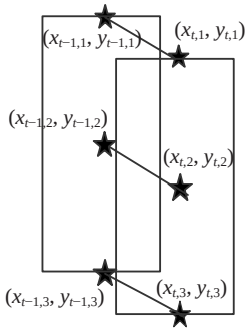



**Fig.5 Median coordinate correspondence in succession**

The same person between two successive frames has the nature of a larger probability value in the course of the movement of people. From Fig.6, we can see that the median coordinates of a previous person A, person B, person C at time $t-1$ are subtracted from median coordinates of current person A at time $t$. For person A at time $t$, we can acquire three motion density functions $P_{AA'}(\boldsymbol{v}_t)$, $P_{AB'}(\boldsymbol{v}_t)$, and $P_{AC'}(\boldsymbol{v}_t)$. Finally, the value of $P_{AA'}(\boldsymbol{v}_t)$ is the largest one among $P_{AA'}(v_t)$, $P_{AB'}(v_t)$, and $P_{AC'}(v_t)$ in general. For person B at time $t$, the value of $P_{BB'}(\boldsymbol{v}_t)$ is the largest one among $P_{BA'}(\boldsymbol{v}_t)$, $P_{BB'}(\boldsymbol{v}_t)$ and $P_{BC'}(\boldsymbol{v}_t)$ in general. For person C the value of $P_{CC'}(\boldsymbol{v}_t)$ is the largest one among $P_{CA'}(\boldsymbol{v}_t)$, $P_{CB'}(\boldsymbol{v}_t)$ and $P_{CC'}(\boldsymbol{v}_t)$ in general. It will be seen from this that the motion density function value of the same person between two successive frames is much larger.

When two persons walk closely, the motion model is able to work. However, in this case the two persons will have the approximate probability value by motion density function. So the contribution of the color model is much greater than that of the motion model. Thus, the occlusion happens, the tracking mainly depending on the color model.

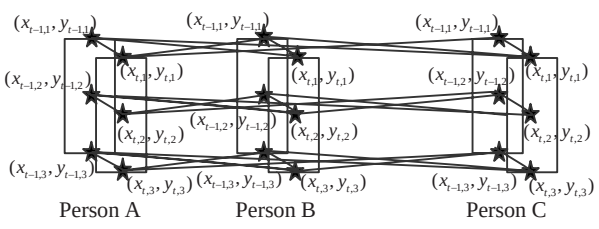



**Fig.6 Motion model for many people**

## TRACKING PEOPLE UNDER OCCLUSION AND ACROSS CAMERAS

We make use of space, motion and color information in the course of tracking multiple people accurately in densely crowded scenes. We will use the appearance model and motion model to solve tracking people on one camera. Given a stored model of a person $k$ at time $t-1$ $M_{t-1,k}=\{B_1, B_2, B_3, B_4, \ldots\}$ to which we referred in Section 2, and the median coordinate of each foreground region $\boldsymbol{v}_t=(x_{t,1}, y_{t,1}, x_{t,2}, y_{t,2}, x_{t,3}, y_{t,3})$, the probability $P(\boldsymbol{c}_{t,x,y}, \boldsymbol{v}_t|M_{t-1,k})$ is equal to

$$\begin{aligned} P(\boldsymbol{c}_{t,x,y}, \boldsymbol{v}_t \mid M_{t-1,k}) &= P(\boldsymbol{c}_{t,x,y} \mid M_{t-1,k}) P(\boldsymbol{v}_t \mid M_{t-1,k}) \\ &= \max\{P(\boldsymbol{c}_{t,x,y} \mid B_1), P(\boldsymbol{c}_{t,x,y} \mid B_2), P(\boldsymbol{c}_{t,x,y} \mid B_3), \ldots\} \\ &\quad \cdot P(\boldsymbol{v}_t \mid M_{t-1,k}) \\ &= \max\{P_{B_1}(\boldsymbol{c}_{t,x,y}), P_{B_2}(\boldsymbol{c}_{t,x,y}), P_{B_3}(\boldsymbol{c}_{t,x,y}), \ldots\} P_{M_{t-1,k}}(\boldsymbol{v}_t). \end{aligned} \tag{12}$$

We consider the likelihood maximization in all the subsets $M_{t-1,k}=\{B_1, B_2, B_3, B_4, \ldots\}$as the probability of $\boldsymbol{c}_{t,x,y}$, $P(\boldsymbol{c}_{t,x,y}, \boldsymbol{v}_t|M_{t-1,k})$. The greater the probability, the closer it will correspond to the blob.

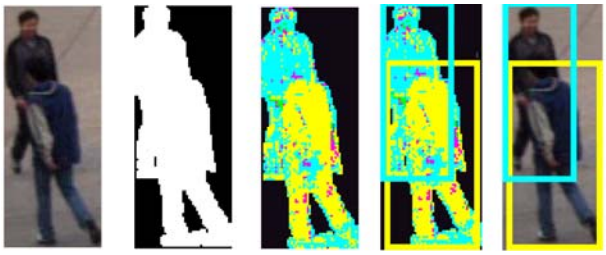
So far, we have achieved the probability $P(\boldsymbol{c}_{t,x,y}, \boldsymbol{v}_t|M_{t-1,k})$ relative to a certain person at time $t-1$. Given a sample foreground set of one person $S=\{\boldsymbol{x}_i\}$, $i=1, 2, \ldots, N$ at time $t$ and a median coordinate of each foreground region $\boldsymbol{v}_t=(x_{t,1}, y_{t,1}, x_{t,2}, y_{t,2}, x_{t,3}, y_{t,3})$, the current pixel $\boldsymbol{c}_{t,x,y}$ which belongs to $S$ can be classified into one of the person models as follows:

$$\begin{aligned} &\boldsymbol{c}_{t,x,y} \in S \quad \text{s.t.} \quad k = \arg_k \max P(\boldsymbol{c}_{t,x,y}, \boldsymbol{v}_t \mid M_{t-1,k}), \\ &\qquad\qquad k = 1, 2, \ldots, n, \end{aligned} \tag{13}$$

where $n$ represents the count of people in the stored model in one camera, $S$ stands for a foreground pixel set of the current one person, $\boldsymbol{c}_{t,x,y}$ stands for one pixel of set $S$, and $M_{t-1,k}$ represents the $k$th person model in the stored model at last frame $t-1$. The meaning of the formula is that a certain person in the stored model will be found who is the most similar to the current pixel $\boldsymbol{c}_{t,x,y}$ in the current foreground pixel set $S$. A pixel $\boldsymbol{c}_{t,x,y}$ can be classified to one of the stored models using maximum likelihood. Finally, the set of $S$ can be divided into many subsets, in which some of the largest subsets will determine the property of the

person of set *S*.

Firstly we reconstruct the background of the video. Then we detect foreground pixels using background subtraction and remove the shadow (Fig.7b). And then we compute probability $P(\boldsymbol{c}_{t,x,y}, \boldsymbol{v}_t|M_{t-1,k})$. Each pixel is correspondent to a certain color according to Eq.(13). Fig.7c is the likelihood maximization image. One person is composed of pixels of the same color on the whole. Finally, we segment occlusions according to the same color which represents one person (Fig.7d). Fig.7e is the final result.



(a) (b) (c) (d) (e)

**Fig.7 Processes of the tracking algorithm**

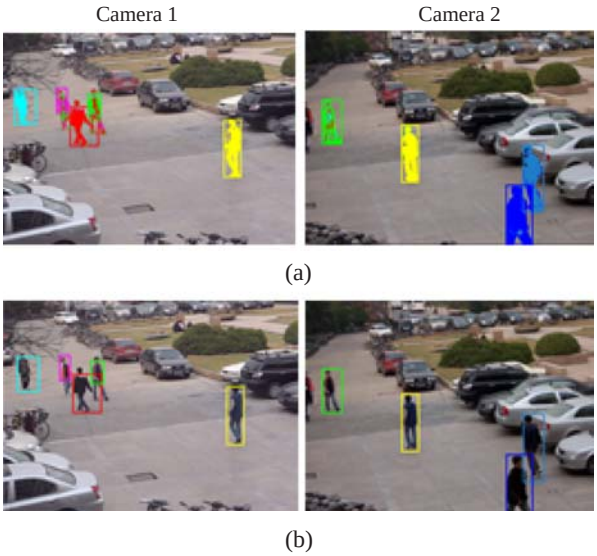
(a) Original image; (b) Detected foreground image; (c) Maximization likelihood image; (d) Segmenting occlusion according to color image; (e) Segmenting result image

This algorithm is different from mean-shift tracking. In mean-shift tracking the blob to be tracked is placed on the current image and a weight for each pixel is computed. The weights are simply the ratio of the probabilities of the estimated histogram, etc. As mentioned above, the contribution of this study is the use of the color density function and motion density function on the basis of the cluster, by means of the multivariate kernel density estimation.

In this part we will apply the appearance model to the tracking of people across cameras. As we know, our cameras are used to extend the field of view with overlapping field of view, which are distinguished from the same field of view (Fleuret *et al*., 2007) and non-overlapping field of view (Javed *et al*., 2003). We do not use homography in correspondences between cameras. The algorithm for tracking people across cameras is different from that for tracking people on one camera. The motion model cannot be utilized, so Eq.(l3) can be changed as

$$\boldsymbol{c}_{t,x,y} \in S \text{ s.t. } k = \arg_k \max P(\boldsymbol{c}_{t,x,y} \mid M_{t-1,k}),$$
$$k = 1, 2, \ldots, n, \qquad (14)$$

where $P(\boldsymbol{c}_{t,x,y}|M_{t-1,k})$ is the appearance model. Similarly, the set of *S* can be divided into many subsets, in which the largest subset will determine the nature of the person in set *S*. We can implement the method to match persons between cameras by the maximum likelihood image in Fig.8a, which is computed by Eq.(14).




(a)

(b)

**Fig.8 Tracking people across camera 1 and camera 2**

(a) Maximum likelihood images; (b) Tracking result across cameras

From Fig.8, we can see that two cameras have a common field of view. Hence, when one person (such as the leftmost person in camera 2 in Fig.8) enters the second view, the stored model of the person must exist in camera 1. We match the person in camera 2 with the stored model in camera 1 according to Eq.(14). When the person gets matched, we will construct a new stored model for this person in camera 2. As described above, our algorithm can address the problem of keeping track of people across cameras. However, the algorithm of mean-shift tracking can only be used in the case of single cameras.

## EXPERIMENTAL RESULTS

We have implemented the proposed method in C++ and found it to work very robustly. Our video sequences are captured from outdoor environments with two fixed cameras. In our experiments, a tracked person is represented with a gray bounding box. Different persons are labeled with different grays. For

the proposed algorithm, people are correctly detected, matched and tracked in two fields of view in most cases, except the following three cases:

(1) Two persons in the group are both dressed in a similar color;

(2) People in the image are too small to be detected;

(3) Some persons in a group enter the field of view from the left hand side of camera 1 or some persons under occlusions enter the field of view from the right hand side of camera 2. In this case, models of people cannot be acquired beforehand.

**Results of segmenting occlusions**

In Fig.9, we show the processes and results of the tracking algorithms which are used to solve occlusions of camera 1 sequences and camera 2 sequences. At first we reconstruct the background of the video. Then we detect foreground pixels using background subtraction and remove the shadow (Fig.9b). And then we compute the likelihood probabilities. Each likelihood probability is correspondent to a color. The correspondent color to the largest likelihood probability is the final color for each blob in the current frame. From Fig.9c, we can see that the same colors make up one person, on the whole. Finally, we segment occlusions according to the same color which represents one person (Fig.9d). As shown in Fig.9d, we get better results.

**Correspondences results between cameras**

In Fig.10a, from the top to the bottom, the frame numbers are 453 and 543, respectively. In this experiment, five persons enter the field from camera 1 to camera 2 in succession and three persons enter the field from camera 2 to camera 1 in succession. They will meet one another in the two fields; that is, the conditions under occlusions and across cameras are created. At Frame 453, one person (the pink bounding box in color space) enters the second view and the method across cameras can track the person from camera 1, so this person in camera 2 has the same color as the person in camera 1. When the new person comes into the field of view from the left hand side of camera 1 or from the right hand side of camera 2, we will construct a model for this person. At Frame 453, there is a person who does not enter the field of view in full. So we will construct a model for him when he enters the field of view in full.

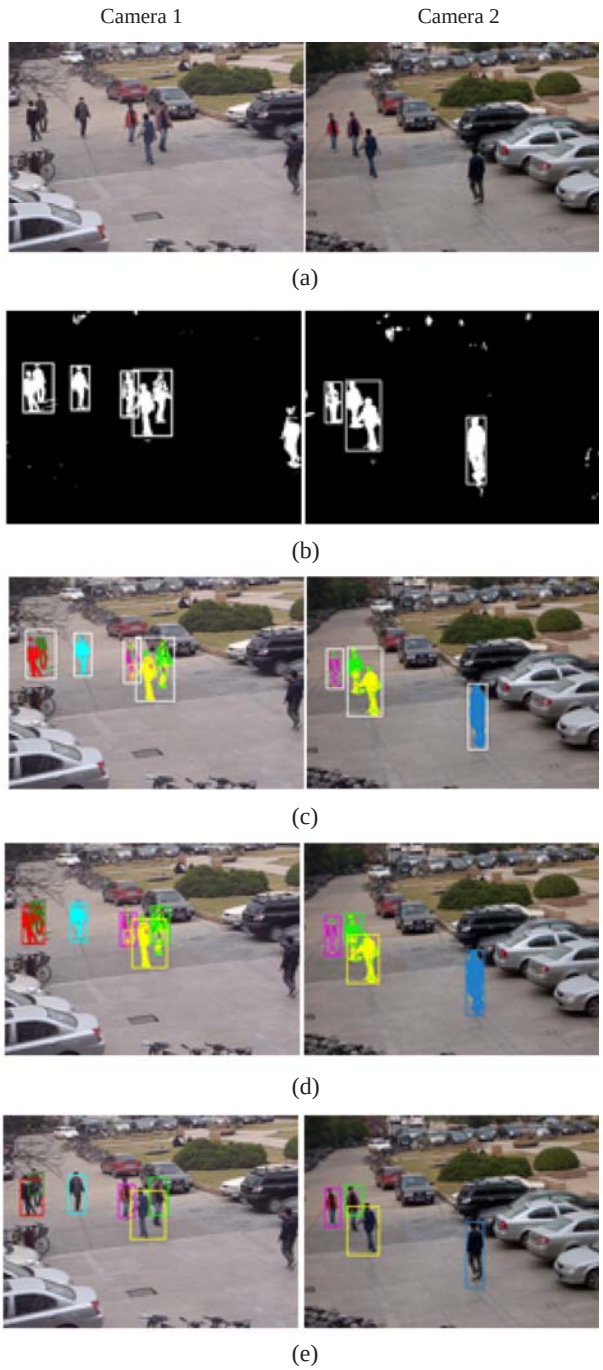



**Fig.9 Processes of the tracking algorithm of camera 1 and camera 2**

(a) Original images; (b) Detected foreground images; (c) Likelihood maximization images (one person labeled one color); (d) Segment occlusions according to color image; (e) Segment result images

We also apply the proposed method to another video to declare the superiority of the method. In Fig.10b, there are two persons who will leave the field of view. So the stored models of them are removed. Our methods can also solve the correspondences between cameras in the case of many occlusions (Fig.10).

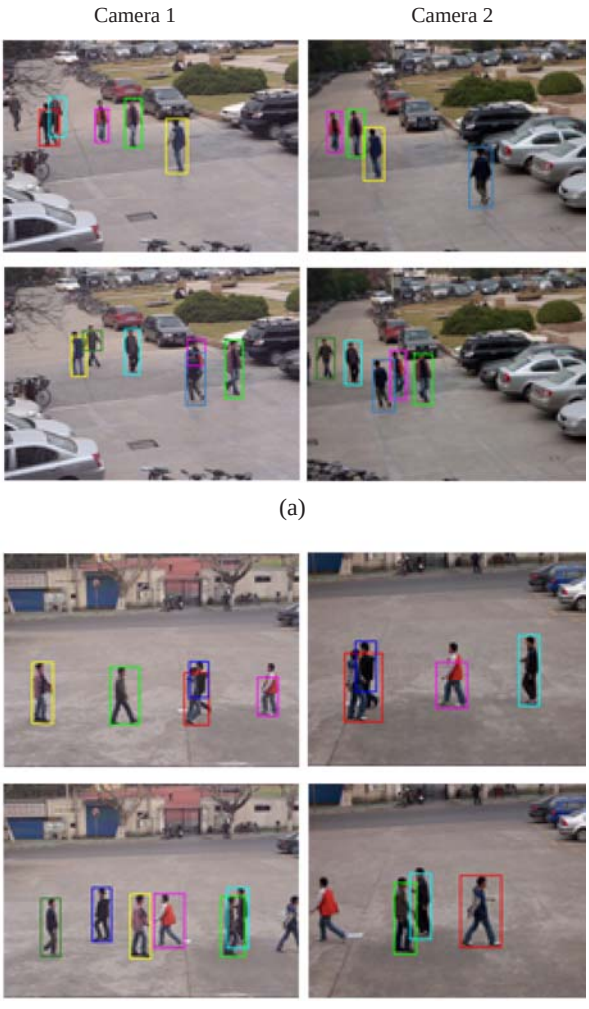



**Fig.10 Tracking and correspondence of multiple people with cameras. (a) Video A; (b) Video B**

The acquired trajectories of eight persons in video A and seven persons in video B, each video including two views, are shown in the reconstruction images and the *X*-*Y* plane in Fig.11. Some isolated singularities are caused by the occlusions; i.e., just part of the body can be seen of the inner person (The person of blue trajectory enters the left view from between cars in video A. He is occluded by the cars. So the trajectory is on the car. Because the trajectories of some people are too dense, we plot the traces on the ground for each person in every two frames).

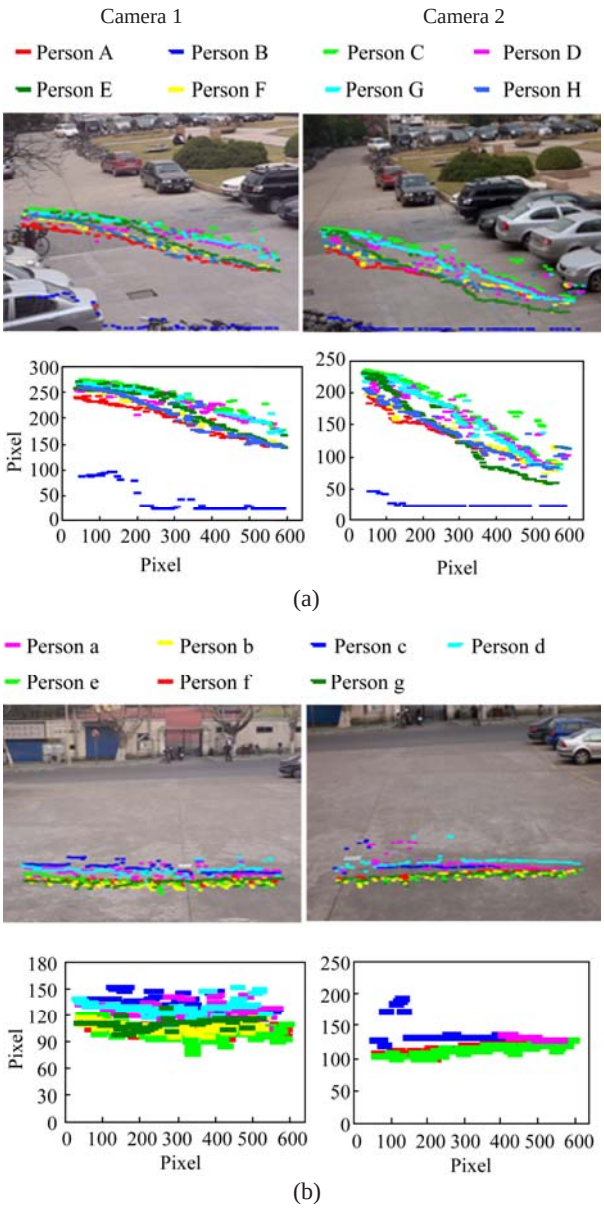



**Fig.11 Trajectories of persons tracked**
(a) The trajectories in real image and the *X*-*Y* plane with video A;
(b) The trajectories in real image and the *X*-*Y* plane with video B

In order to evaluate the algorithms, we compare the tracking valid times of each person with true appearance times which are obtained manually. Fig.12 shows the comparison results of video A and video B, where the horizontal axis represents the persons and the vertical axis represents the appearance times of each person, including the case of occlusions. From

Fig.12 we can see that the approach in this research obtains a much better effect. In addition, one of these advantages is that we can correct the false into the true in the next frame if one person is tracked falsely in the previous frame.

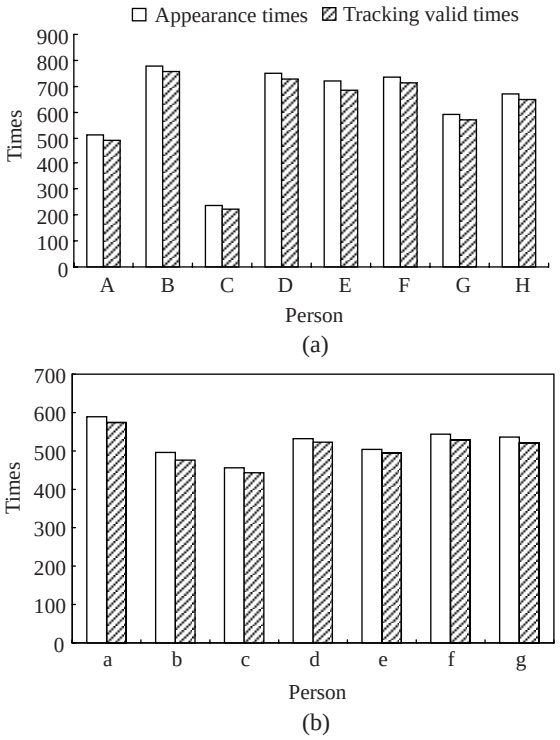



**Fig.12 Comparison between the appearance times and the tracking valid times. (a) Video A ; (b) Video B**

## CONCLUSION AND DISCUSSION

We can summarize our contribution as follows: We address the problems of how to keep track of multiple people under occlusion and across cameras with overlapping field of view using a probabilistic model which includes a blob model, color model and motion model. Robust performance is achieved through the integration of three key models. The proposed algorithms do not need camera calibration. Unlike many systems for tracking people, we make no use of homography cues. Instead, we employ a combination of space, color and motion information in a meaningful way. Our algorithm in this study is an automation tracking algorithm in the whole course as opposed to a mean-shift tracking algorithm. Our algorithm implements the cluster using the advantage of gradient optimization of the mean shift and optimizes space information and color information. We track the people under occlusions and across cameras by means of the maximization likelihood image, which is obtained by the color density function and motion density function on the basis of the cluster. This algorithm is capable of simultaneously tracking multiple people even with occlusion.

As far as multiple cameras are concerned, they are used to extend the field of view in this study, because one camera has a limited field of view. So they are distinguished from multiple cameras which have a common field of view. People correspondence across cameras in this study is based on a color model. This novel color model does not restrict the clothing to be of uniform color. Instead, it can be any mixture of colors. The experimental results have demonstrated the effectiveness and robustness of our method.

Future work includes segmenting groups of people in the event that they are occlusions on going into views. The dynamics/kinematics model of human motion and Kalman filters (Wren and Pentland, 1998) will be utilized to help the tracking process.

## References

Angel, D.S., Aifanti, N., Malassiotis, S., Strintzis, M.G., 2005. Prior knowledge based motion model representation. *Electr. Lett. Comput. Vis. Image Anal.*, **5**(3):55-67.

Black, J., Ellis, T., Rosin, P., 2002. Multi-view Image Surveillance and Tracking. IEEE Workshop on Motion and Video Computing, p.169-174. [doi:10.1109/MOTION.2002.1182230]

Comaniciu, D., Meer, P., 1999. Mean Shift: analysis and Applications. IEEE Int. Conf. on Computer Vision, Kerkyra, Greece, p.1197-1203.

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**(5):603-619. [doi:10.1109/34.1000236]

Cucchiara, R., Grana, C., Piccardi, M., Prati, A., 2003. Detecting moving objects, ghosts, and shadows in video streams. *IEEE. Trans. Pattern Anal. Mach. Intell.*, **25**(10):1337-1342. [doi:10.1109/TPAMI.2003.1233909]

Du, W., Piater, J., 2007. Multi-camera People Tracking by Collaborative Particle Filters and Principal Axis-based Integration. Asian Conf. on Computer Vision, p.365-374.

Duong, T., Hazelton, M.L., 2005. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinav. J. Statist.*, **32**(3):485-506. [doi:10.1111/j.1467-9469.2005.00445.x]

Fleuret, F., Berclaz, J., Lengagne, R., Fua, P., 2007. Multi-camera people tracking with a probabilistic occupancy

map. *IEEE. Trans. Pattern Anal. Mach. Intell.*, **30**(2): 267-282. [doi:10.1109/TPAMI.2007.1174]

Giné, E., Koltchinskii, V., Zinn, J., 2004. Weighted uniform consistency of kernel density estimators. *Inst. Math. Stat. Ann. Probab.*, **32**(3B):2570-2605.

Han, M., Xu, W., Tao, H., Gong, Y., 2004. An Algorithm for Multiple Object Trajectory Tracking. Conf. on Computer Vision and Pattern Recognition, **1**:864-871.

Isard, M., MacCormick, J., 2001. Bramble: A Bayesian Multiple-blob Tracker. Conf. on Computer Vision and Pattern Recognition, **2**:34-41.

Javed, O., Rasheed, Z., Shafique, K., Shah, M., 2003. Tracking Across Multiple Cameras with Disjoint Views. Proc. 9th IEEE Int. Conf. on Computer Vision, **2**:952-957. [doi:10.1109/ICCV.2003.1238451]

Kang, J., Cohen, I., Medioni, G., 2004. Tracking People in Crowded Scenes Across Multiple Cameras. Asian Conf. on Computer Vision.

Khan, S., Shah, M., 2000. Tracking People in Presence of Occlusion. Asian Conf. on Computer Vision.

Khan, S., Shah, M., 2003. Consistent labeling of tracked objects in multiple cameras with overlapping field of view. *IEEE Trans. Pattern Anal. Mach. Intell.*, **25**(10):1355-1360. [doi:10.1109/TPAMI.2003.1233912]

Khan, S., Shah, M., 2006. A Multiview Approach to Tracking People in Crowded Scenes Using a Planar Homography Constraint. Proc. European Conf. on Computer Vision.

Lim, H., Morariu, V.I., Camps, O.I., Sznaier, M., 2006. Dynamic Appearance Modeling for Human Tracking. Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition.

Mittal, A., Davis, L.S., 2003. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. *Int. J. Comput. Vis.*, **51**(3):189-203. [doi:10.1023/A:1021849801764]

Otsuka, K., Mukawa, N., 2004. Multi-view Occlusion Analysis for Tracking Densely Populated Objects Based on 2-D Visual Angles. Conf. on Computer Vision and Pattern Recognition.

Romano, R., Lee, L., Stein, G., 2000. Monitoring activities from multiple video streams: Establishing a common coordinate frame. *IEEE Trans. PAMI*, **22**(8):758-768.

Scott, D.W., Sain, S.R., 2004. Multi-dimensional Density Estimation. Elsevier Science.

Smith, K., Gatica-Perez, D., Odobez, J.M., 2005. Using Particles to Track Varying Numbers of Interacting People. Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, **1**:962-969.

Stauffer, C., Grimson, W.E.L., 1999. Adaptive Background Mixture Models for Real-time Tracking. Proc. IEEE Conf. on Computer Vision and Pattern Recognition.

Sullivan, J., Blake, A., Isard, M., MacCormick, J., 1999. Object Localization by Bayesian Correlation. Proc. 7th Int. Conf. on Computer Vision, **2**:1068-1075.

Vega, I.R., Sarkar, S., 2003. Statistical motion model based on the change of feature relation ships: human gait-based recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, **25**(10):1323-1328. [doi:10.1109/TPAMI.2003.1233906]

Wren, C.R., Pentland, A.P., 1998. Dynamic Modeling of Human Motion. Proc. 3rd IEEE Int. Conf. on Automatic Face and Gesture Recognition, p.22-27.

Yu, Y., Harwood, D., Yoon, K., Davis, L.S., 2007. Human appearance modeling for matching across video sequences. *Mach. Vis. Appl.*, **18**(3-4):139-149. [doi:10.1007/s00138-006-0061-z]