Foreground Object Detection from Videos Containing Complex Background

Liyuan Li, Weimin Huang
Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore, 119613
lyli,wmhuang@i2r.a-star.edu.sg

Irene Y.H. Gu
Dept. Signals & Systems
Chalmers Univ. of Tech.
41296, Sweden
irenegu@s2.chalmers.se

Qi Tian
Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore, 119613
tian@i2r.a-star.edu.sg

ABSTRACT

This paper proposes a novel method for detection and segmentation of foreground objects from a video which contains both stationary and moving background objects and undergoes both gradual and sudden "once-off" changes. A Bayes decision rule for classification of background and foreground from selected feature vectors is formulated. Under this rule, different types of background objects will be classified from foreground objects by choosing a proper feature vector. The stationary background object is described by the color feature, and the moving background object is represented by the color co-occurrence feature. Foreground objects are extracted by fusing the classification results from both stationary and moving pixels. Learning strategies for the gradual and sudden "once-off" background changes are proposed to adapt to various changes in background through the video. The convergence of the learning process is proved and a formula to select a proper learning rate is also derived. Experiments have shown promising results in extracting foreground objects from many complex backgrounds including wavering tree branches, flickering screens and water surfaces, moving escalators, opening and closing doors, switching lights and shadows of moving objects.

Categories and Subject Descriptors

I.4 [**Image Processing And Computer Vision**]: Segmentation—*pixel classification*

General Terms

Algorithms

Keywords

Video processing, background modeling, foreground segmentation, video surveillance, Bayes model, color co-occurrence.

MM'03, November 2–8, 2003, Berkeley, California, USA.

1. INTRODUCTION

Foreground object detection and segmentation from a video stream is one of the essential tasks in video processing, understanding, and object-based video encoding (e.g., MPEG4). An commonly used approach to extract foreground objects from the image sequence is through background suppression, or background subtraction and its variants [3, 11, 12] when the video is grabbed from a stationary camera. These techniques have been widely used in real-time video processing. However, the task becomes difficult when the background contains shadows and moving objects, e.g., wavering tree branches and moving escalators, and undergoes various changes, such as illumination changes and moved objects.

Many methods have been proposed for real-time foreground object detection from video sequences. However, most of them were developed under the assumption that the background consists of stationary objects whose color or intensity may change gradually over time. The simplest way is to smooth the color of a background pixel with an Infinite Impulse Response (IIR) or a Kalman filter [5, 6] through real-time video. A better way to tolerate the background variation in the video is to employ a Gaussian function that describes the color distribution of each pixel belonging to a stable background object [14, 1]. The Gaussian model parameters are recursively updated in order to follow the gradual background changes in the video.

Recently, several other methods suitable for a variety of background situations have been proposed. Among them, Mixture of Gaussians (MoG) [11, 9] is considered as a promising method. In MoG, the colors from a pixel in a background object are described by multiple Gaussian distributions. Good foreground object detection results were reported by applying MoG to outdoor scenes. Further investigations showed that MoG with more than two Gaussians can degrade the performance in foreground object detection [1, 2]. The background variation model employed in W^4 [3] is a generalization of the Gaussian model. In [12], Toyama *et al* employed a linear Wiener filter to learn and predict color changes in each background pixel through the video. The linear predictor can model both stationary and moving background objects. The weakness of this method is that it is difficult to model the non-periodical background changes. These methods can work for a real-time video containing a variety of background variations. However, they are still difficult to handle a wide range of changes in moving back-

ground objects, e.g., moving background objects and various shadows.

Another way to separate foreground objects from moving background objects is to exploit the consistency of optical flows over a short period of time [13, 4]. The methods were reported as being able to detect foreground objects in complex outdoor scenes that contain nonstationary vegetation. The difficulty of this technique is that computing optical flow is an ill-posed problem at the regions of less texture features and the boundaries of image discontinuities [13]. In our previous work [7], a method has been proposed which employed the color co-occurrence to describe the moving background. The preliminary result indicates that the feature is more effective to model the dynamic parts of the background than those employed by previous methods. However, it can not recover the background from sudden "once-off" environment changes in the video. Hence, a general background model which can incorporate different features is required for complex environments.

In this paper, we propose a novel method to extract foreground objects from a real-time complex video under the Bayes decision framework. A Bayes decision rule for classification of background and foreground from a general feature vector is formulated. Meanwhile, an effective data structure to learn and maintain the statistics of different feature vectors is established. Based on these, two types of features are employed to model the complex background containing both stationary and motion objects. The statistics of most significant colors are used to describe the stationary parts of the background, and that of most significant color co-occurrences are used to describe the motion objects of the background. Foreground objects are extracted by fusing the detection results from both stationary and motion points. Meanwhile, learning strategies for gradual and "once-off" background changes are proposed. Compared with our previous work in [7], several new extensions are introduced. First, the Bayes decision rule has been extended to general features. Under this framework, multiple features can be integrated for background and foreground classification. Hence, the proposed method can not only model the motion background objects but also deal with sudden "once-off" changes and multiple states in the stationary background. Secondly, mathematical proof about the convergence of the learning process is given. Finally, more extensive experiments on difficult situations and some quantitative evaluations have been reported. Many tested videos contain complex background objects and, in our knowledge, no existing methods had been tested on such complex cases before this work.

The remaining of the paper is organized as follows. In Section 2, we first give a specification about the background scene and background changes in videos. Then we formulate the background and foreground classification problem by utilizing a general feature vector based on Bayes theory. The data structure to learn and maintain the statistics for different types of background features is established. It also describes how the features are selected for a complex background. Section 3 describes the algorithm for foreground object segmentation based on background modeling and Bayes classification. It contains four parts: change detection, background and foreground classification, foreground object segmentation, and background learning and maintenance. The experimental results on various complex videos and the quantitative evaluation are presented in Section 4. The paper is concluded in Section 5.

2. BAYES CLASSIFICATION OF BACKGROUND AND FOREGROUND

2.1 Problem Specification

For the general purpose of video processing, the background is usually considered as the scene without the presence of objects of interest, such as human objects or moving vehicles. Background is usually composed of non-living objects that remain passively in the scene. In a video about a general environment, the background can consist of both stationary and moving objects. The stationary background objects can be walls, doors, and furniture in an indoor scene, as well as buildings, vegetation, and ground surfaces in an outdoor scene. The moving background objects can be wavering tree branches, flickering water surfaces or screens of computers, wavering curtains, moving fans, running escalators, and many more. Meanwhile, the background might be undergoing two types of changes over the time. One is the gradual changes caused by natural lighting variations, e.g., the change of illumination from day to night. The other is the sudden "once-off" changes. The global sudden "once-off" changes may be caused by switching on/off some lights or the change of view angle of a camera, and the local "once-off" changes may be caused by removing or depositing the background objects, e.g., moving a chair to a different position. Besides, the foreground object might be converted to be a background object, such as a car moving into a parking lot. In some cases, a background pixel may have multiple states, such as sunny and cloudy scenes. Therefore, for a complex environment, different parts of the background should be described with different types of features. However, almost all existing methods only employ one type of features, e.g., color or optical flow, to model both static and dynamic parts of the background. In this paper, we propose a general Bayesian framework which can integrate multiple features to model the background for foreground object detection.

2.2 Formulation of the Classification Rule

For one type of background object, there exist some significant features that can be exploited to effectively separate the background from the foreground objects. Let $\mathbf{v}_t$ be a discrete value feature vector extracted from an image sequence at the pixel $s = (x, y)$ and time instant t. Using Bayes rule, it follows that the a posterior probability of $\mathbf{v}_t$ from the background b or foreground f is

$$P(C|\mathbf{v}_t, s) = \frac{P(\mathbf{v}_t|C, s)P(C|s)}{P(\mathbf{v}_t|s)}, \quad C = b \text{ or } f \qquad (1)$$

Using the Bayes decision rule, the pixel is classified as background if the feature vector satisfies

$$P(b|\mathbf{v}_t, s) > P(f|\mathbf{v}_t, s) \qquad (2)$$

Noting that the feature vectors associated the pixel s are either from background or from foreground objects, it follows

$$P(\mathbf{v}_t|s) = P(\mathbf{v}_t|b, s) \cdot P(b|s) + P(\mathbf{v}_t|f, s) \cdot P(f|s). \qquad (3)$$

Substituting (1) and (3) to (2), it becomes

$$2P(\mathbf{v}_t|b, s) \cdot P(b|s) > P(\mathbf{v}_t|s) \qquad (4)$$

This indicates that by learning the a prior probability $P(b|s)$, the probability $P(\mathbf{v}_t|s)$ and the conditional probability $P(\mathbf{v}_t|b,s)$ in advance, we may classify a feature $\mathbf{v}_t$ as either associated with foreground or with background.

2.3 Representation of Feature Statistics

The mathematical form of $P(\mathbf{v}_t|s)$ and $P(\mathbf{v}_t|b,s)$ in (4) are unknown in general cases. They could be represented by the histograms of feature vectors over the entire feature space. For a n dimensional feature vector with L quantization levels, the joint histogram for $P(\mathbf{v}_t|s)$, or $P(\mathbf{v}_t|b,s)$ contains L^n bins. If L or n is large, operating on the joint histogram would be expensive both for computation and storage. A good approximation is therefore desirable.

Background is considered as containing non-living objects which stay constantly in the same place in the scene, while objects of interest would often move in the scene. Hence, if the selected features are effective to represent background, at a pixel s, the feature vectors from the background would concentrate in a very small subspace of the feature histogram, while the feature vectors from foreground objects would distribute widely in the feature space. This means, with a good feature selection, it becomes possible to cover a large percentage (e.g., more than 90%) of the feature vectors associated with the background by using a small number of bins in the histogram. Meanwhile, less feature vectors from foreground objects would be distributed in these few bins.

Let $P(\mathbf{v}_t^i|b,s)$, $i = 1,\cdots,N$, be the first N bins from the histogram sorted according to the descendent order of $P(\mathbf{v}_t|b,s)$, i.e., $P(\mathbf{v}_t^i|b,s) \geq P(\mathbf{v}_t^{i+1}|b,s)$. For giving percentage values M_1 and M_2, e.g, $M_1 = 90\%$ and $M_2 = 10\%$, there exists a small integer N_1 such that the following conditions are satisfied

$$\sum_{i=1}^{N_1} P(\mathbf{v}_t^i|b,s) > M_1 \quad \text{and} \quad \sum_{i=1}^{N_1} P(\mathbf{v}_t^i|f,s) < M_2 \qquad (5)$$

Naturally, N_1 value is also dependent on the selected features and the number of quantitative levels used for the features.

For each type of feature vectors (either from foreground or background), *a table of feature statistics*, denoted as $S_{\mathbf{v}_t}^{s,t,i}$, $i = 1,\cdots,N_2$ ($N_2 > N_1$), is maintained at pixel s and time t to record the statistics for the N_2 most significant values. Each element in the table consists of three components, i.e.,

$$S_{\mathbf{v}_t}^{s,t,i} = \begin{cases} p_v^{t,i} = P(\mathbf{v}_t^i|s) \\ p_{vb}^{t,i} = P(\mathbf{v}_t^i|b,s) \\ \mathbf{v}_t^i = [a_1^i,\cdots,a_n^i]^T \end{cases} \qquad (6)$$

The elements in the list are sorted according to the descendent order of $p_v^{t,i}$. Meanwhile, $p_b^{s,t} = P(b|s)$ is also maintained at each time t to adapt to the variation of busyness in the scene over different time duration. The list forms the most significant portion of the histogram for the feature vector $\mathbf{v}_t$. For many video processing applications, the duration of background exposure at a pixel is much longer than the time being covered by foreground objects. The first N_1 elements of the list is enough to cover the majority part of the feature vectors from the background. Hence, the first N_1 elements from the table together with $p_b^{s,t}$ are used for the classification of background and foreground changes. Within the first N_1 elements in the table, there are $p_v^{t,i} \approx p_{vb}^{t,i}$ for the background features, otherwise, $p_v^{t,i} \gg p_{vb}^{t,i}$. The elements from N_1 to N_2 in the list are used as a buffer to learn the new significant features through the background updating. The values of N_1 and N_2 are selected empirically. For the stable features from the background, small values are good enough. For the features with variations, slightly large values are required.

2.4 Selection of Feature Vectors

When a pixel is associated with a stationary background object, the colors of it are naturally selected as feature vectors, i.e., $\mathbf{v}_t$ in (1) to (6) is substituted by $\mathbf{c}_t = [r_t\ g_t\ b_t]^T$. When a pixel is associated with a moving background object, the color co-occurrences of inter-frame changes from the pixel are chosen as the feature vectors, i.e., $\mathbf{v}_t$ in (1) to (6) is substituted by $\mathbf{cc}_t = [r_{t-1}\ g_{t-1}\ b_{t-1}\ r_t\ g_t\ b_t]^T$. The selection of color co-occurrences are based on the following observation. For a moving background object, even though the colors from a pixel associated with it varies greatly, the color co-occurrences of the inter-frame changes from it is quite significant since the similar changes always happen in the same place in the image. To represent multiple states of background at a pixel, such as the moving of tree branch and the exposure of sky, both $S_{\mathbf{c}_t}^{s,t,i}$ and $S_{\mathbf{cc}_t}^{s,t,i}$ are maintained at each pixel.

To make the computation and storage efficiency, $L = 64$ quantization levels in each color component are used for color vector. Meanwhile, $N_1 = 30$ and $N_2 = 50$ are chosen. $L = 32$ with $N_1 = 50$ and $N_2 = 80$ are used for the feature vectors of color co-occurrences. In this investigation, it was found with such parameter selections, (5) holds for most pixels associated with both stationary and moving background objects. Obviously, there are 64^3 and 32^6 bins in the joint histograms for color and color co-occurrence vectors, respectively. With $N_1 = 30$ for color features and $N_1 = 50$ for color co-occurrence features, good representation of background has been achieved for the stationary and moving background objects. This means the selected features are quite effective.

3. ALGORITHM DESCRIPTION

With the formulation of background and foreground classification based on Bayes decision theory, an algorithm for foreground object detection from a real-time video containing complex background are established. It consists of four parts: change detection, change classification, foreground object segmentation, and background learning and maintenance. The block diagram of the proposed algorithm is shown in Figure 1. The light blocks from left to right correspond to the first three steps, and the gray blocks for the step of adaptive background modeling. In the first step, non-change pixels in the image stream are filtered out by using simple background and temporal differences. The detected changes are separated as pixels belonging to stationary and moving objects according to inter-frame changes. In the second step, the pixels associated with stationary or moving objects are further classified as background or foreground based on the learned statistics of colors and color co-occurrences respectively by using the Bayes decision rule. In the third step, foreground objects are segmented by combining the classification results from both stationary and moving parts. In the fourth step, background models are updated. Both gradual and "once-off" learning strategies are utilized to learn the statistics of feature vectors. Meanwhile,

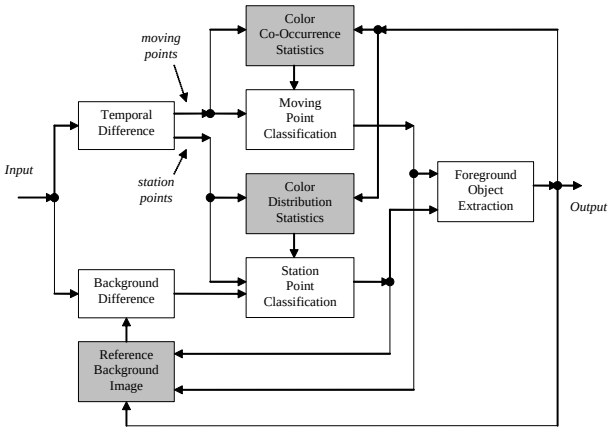

Figure 1: The block diagram of the algorithm.

a reference background image is maintained to make the background difference accurate and adaptive to the changing background. The details for the four steps are described in the following subsections.

3.1 Change Detection

In the first step, pixels of insignificant changes are filtered out by simple background and temporal differencing. Let $I(s,t) = \{I_c(s,t)\}$ be the input color image and $B(s,t) = \{B_c(s,t)\}$ be the reference background image maintained by the system at time instant t, and $c \in \{r, g, b\}$ represents a color component. The background and temporal differencing are performed as follows. First, a simple picture differencing is performed for each color component with adaptive thresholding, using the method described in [10]. The results from the three components are combined to generate the background difference $F_{bd}(s,t)$ and the temporal difference $F_{td}(s,t)$ respectively. The image differencing is used to removed the imaging noise. The remainder changes will be classified based on background features. The adaptive global thresholding is accurate for this purpose.

3.2 Change Classification

The temporal differences classify the change pixels into two types. If $F_{td}(s,t) = 1$ is detected, the pixel is classified as a motion pixel belonging to a moving object. Otherwise, it is a station pixel associated with a stationary object. They are further classified as background or foreground separately. For a station pixel s, the color feature vector $\mathbf{v}_t = \mathbf{c}_t = [r_t\ g_t\ b_t]^T$ is generated with $L = 64$ levels for each color component. For a motion pixel, the feature vector of color co-occurrence $\mathbf{v}_t = \mathbf{cc}_t = [r_{t-1}\ g_{t-1}\ b_{t-1}\ r_t\ g_t\ b_t]^T$ is generated with $L = 32$. This feature vector $\mathbf{v}_t$ is then compared with the first N_1 learned feature vectors from the corresponding *table of feature statistics* for background to retrieve the probabilities for the similar features. Let $\mathbf{v}_t = [a_0, \cdots, a_n]^T$ and $\mathbf{v}_t^i$ from the table $S_{\mathbf{v}_t}^{s,t,i}$ (6). The conditional probabilities are obtained as

$$\begin{cases} P(b|s) = p_b^{s,t} \\ P(\mathbf{v}_t|s) = \sum_{j \in M(\mathbf{v}_t)} p_v^{s,t,j} \\ P(\mathbf{v}_t|b,s) = \sum_{j \in M(\mathbf{v}_t)} p_{vb}^{s,t,j} \end{cases} \quad (7)$$

where the matched feature set in $S_{\mathbf{v}_t}^{s,t,i}$ is defined as

$$M(\mathbf{v}_t) = \{k : \forall m \in \{1, \cdots, n\}, |a_m - a_m^k| \le \delta\} \quad (8)$$

where $\delta = 2$ was chosen so that if the similar features are quantized into neighboring vectors, the statistics can still be retrieved. If no element in the table $S_{\mathbf{v}_t}^{s,t,i}$ matches $\mathbf{v}_t$, both $P(\mathbf{v}_t|s)$ and $P(\mathbf{v}_t|b,s)$ are set 0. Substituting the probabilities in (7) into (4), the point are classified as background or foreground.

3.3 Foreground Object Segmentation

It is observed that, after the background and foreground classification, only a small percentage of the background points are wrongly labeled as foreground ones. And more, the remainders have become isolated points. A morphological operation (a pair of *open* and *close*) is applied to remove the scattered error points and connect the foreground points. The remaining regions are extracted with small ones removed. The segmented foreground objects form a binary output image $O(s,t)$.

3.4 Background Learning and Maintenance

Background maintenance adapts the background models to various background changes over time. In the proposed method, the background maintenance includes two parts, updating the tables of feature statistics and a reference background image.

3.4.1 Updating Tables of Feature Statistics

Two tables of color and color co-occurrence statistics are maintained at each pixel. Two updating strategies are proposed to adapt them to both gradual and "once-off" background changes.

3.4.1.1 Updating to gradual background changes.

Assume the feature vector $\mathbf{v}_t$ is used to classify the pixel s as foreground or background at time t, the statistics of the corresponding features (color or color co-occurrence) is gradually updated by

$$\begin{cases} p_b^{s,t+1} = (1-\alpha_2)p_b^{s,t} + \alpha_2 M_b^{s,t} \\ p_v^{s,t+1,i} = (1-\alpha_2)p_v^{s,t,i} + \alpha_2 M_v^{s,t,i} \\ p_{vb}^{s,t+1,i} = (1-\alpha_2)p_{vb}^{s,t,i} + \alpha_2 (M_b^{s,t} \wedge M_v^{s,t,i}) \end{cases} \quad (9)$$

for $i = 1, \cdots, N_2$, where α_2 is the learning rate which controls the speed of feature learning. The selection of the learning rate α_2 will be addressed late in this section. The boolean values for matching labels are generated as follows. $M_b^{s,t} = 1$ when s is labeled as the background at time t from the feedback of final segmentation, otherwise, $M_b^{s,t} = 0$. $M_v^{s,t,i} = 1$ when $\mathbf{v}_t^i$ of $S_{\mathbf{v}_t}^{s,t,i}$ in (6) matches $\mathbf{v}_t$ best and $M_v^{s,t,i} = 0$ for the remainders.

The above updating equation states the following: If the pixel s is labeled as a background point at time t, $p_b^{s,t+1}$ is slightly increased from $p_b^{s,t}$ due to $M_b^{s,t} = 1$. Further, the probability for the matched feature is also increased due to $M_v^{s,t,i} = 1$. If $M_v^{s,t,i} = 0$, then the statistics for the un-matched features are slightly decreased. If there is no match between $\mathbf{v}_t$ and the elements in the table $S_{\mathbf{v}_t}^{s,t,i}$, the N_2th element in the table is replaced by a new feature vector

$$p_v^{s,t+1,N_2} = \alpha_2, \quad p_{vb}^{s,t+1,N_2} = \alpha_2, \quad \mathbf{v}_t^{N_2} = \mathbf{v}_t. \quad (10)$$

If the pixel s is labeled as a foreground point at time t, $p_b^{s,t+1}$ and $p_{vb}^{s,t+1,i}$ are slightly decreased with $M_b^{s,t} = 0$. However,

for the matched element in $S_{\mathbf{v}_t}^{s,t+1,i}$, $p_v^{s,t+1,i}$ is increased. The updated elements in the table $S_{\mathbf{v}_t}^{s,t+1,i}$ are re-sorted on a descendent order for $p_v^{s,t+1,i}$, so that the table may keep the N_2 most frequent features for the corresponding states.

3.4.1.2 Updating to "once-off" background changes.

When an "once-off" background change has happened, the features of the new background appearance become dominated immediately after the change. From (5) and (3), new background features at s is detected if

$$P(f|s)\sum_{i=1}^{N_1} P(\mathbf{v}_t^i|f,s) > T \tag{11}$$

where T is a percentage value which determines when the new features can be recognized as new background appearance. With a large value of T, the system is stable but slow to response to the "once-off" changes. However, if T is small, the system is easy to learn the frequent foreground features as new background appearances. In our tests, T was set as 90%. The factor $P(f|s)$ prevents updating from a small number of features. From (3) and (6), (11) becomes

$$\sum_{i=1}^{N_1} p_v^{s,t,i} - p_b^{s,t}\sum_{i=1}^{N_1} p_{vb}^{s,t,i} > T \tag{12}$$

Noting $P(f|s) = 1 - P(b|s)$ for each type of feature vectors, the statistics of the features are adjusted as follows once the new background features are discovered at s,

$$\begin{cases} p_b^{s,t+1} = 1 - p_b^{s,t} \\ p_{vb}^{s,t+1,i} = (p_v^{s,t,i} - p_b^{s,t}\cdot p_{vb}^{s,t,i})/p_b^{s,t+1} \end{cases} \tag{13}$$

for $i = 1, \cdots, N_1$. With this "once-off" operation, the observed domination features are converted as the learned background features.

3.4.1.3 Convergence of the learning process.

If the few most significant feature vectors represent the background well, there will be $\sum_{i=1}^{N_1} p_{vb}^{t,s,i} \approx 1$. In addition, if the background features become significant, it is desirable that $\sum_{i=1}^{N_1} p_{vb}^{t,s,i}$ will converge to 1 with the evolution of updating. The updating equation (9) meets such requirement.

Suppose there is $\sum_{i=1}^{N_1} p_{vb}^{t,s,i} = 1$ at time t, and at time $t+1$ the jth element of $S_{\mathbf{v}_t}^{s,t,i}$ matches the feature vector $\mathbf{v}_{t+1}$ of image point $I(s,t+1)$ labeled as background. Then there is

$$\sum_{i=1}^{N_1} p_{vb}^{t+1,s,i} = (1-\alpha_2)\sum_{i=1}^{N_1} p_{vb}^{t,s,i} + \alpha_2(M_b^{t,s} \wedge M_v^{t,s,j}) = 1 \tag{14}$$

This means the sum of the probabilities of the background features keeps 1 by the updating equation (9).

Let's suppose $\sum_{i=1}^{N_1} p_{vb}^{t,s,i} \neq 1$ at time t due to some reasons such as the initialization or the operation of "once-off" learning, and the $\mathbf{v}_t^j$ from the first N_1 elements of $S_{\mathbf{v}_t}^{s,t,i}$ matches $\mathbf{v}_{t+1}$ from $I(s,t+1)$, then we have

$$\sum_{i=1}^{N_1} p_{vb}^{t+1,s,i} = (1-\alpha_2)\sum_{i=1}^{N_1} p_{vb}^{t,s,i} + \alpha_2 \tag{15}$$

From (15) one has

$$\sum_{i=1}^{N_1} p_{vb}^{t+1,s,i} - \sum_{i=1}^{N_1} p_{vb}^{t,s,i} = \alpha_2(1 - \sum_{i=1}^{N_1} p_{vb}^{t,s,i}) \tag{16}$$

If $\sum_{i=1}^{N_1} p_{vb}^{t,s,i} < 1$, there will be $\sum_{i=1}^{N_1} p_{vb}^{t+1,s,i} > \sum_{i=1}^{N_1} p_{vb}^{t,s,i}$. The sum of the probabilities for background features increases slightly. On the other hand, If $\sum_{i=1}^{N_1} p_{vb}^{t,s,i} > 1$, there will be $\sum_{i=1}^{N_1} p_{vb}^{t+1,s,i} < \sum_{i=1}^{N_1} p_{vb}^{t,s,i}$. The sum of the probabilities for background features decreases slightly. From these two cases it can be concluded that the sum of the probabilities for background features converges to 1 as long as the background features are significant and the most frequent.

3.4.1.4 Parameter selection for learning.

Another interesting problem about the learning strategy is the selection of the learning rate α_2. To make the gradual updating operation adapt to the gradual background changes smoothly and not to be perturbed by noise and foreground objects too much, the α_2 should be selected small. On the other hand, if α_2 is too small, the system will become too slow to response the "once-off" background changes. Here, we give a formulation to select α_2 from the required time to response to "once-off" background changes.

Let's examine the response of the learning strategy to an "once-off" background change. An ideal "once-off" background change at time t_0 can be assumed as an step function. The features before t_0 fall into the first K_1 elements of $S_{\mathbf{v}_t}^{s,t,i}$ ($K_1 < N_1$), and the features after t_0 fall into the next K_2 elements of $S_{\mathbf{v}_t}^{s,t,i}$ ($K_2 < N_1$). So at time t_0 there are

$$\begin{array}{l} \sum_{i=1}^{K_1} p_v^{t_0,s,i} \approx \sum_{i=1}^{K_1} p_{vb}^{t_0,s,i} \approx p_{b,v}^{t_0,s} \approx 1, \\ \sum_{i=K_1+1}^{K_1+K_2} p_v^{t_0,s,i} \approx \sum_{i=K_1+1}^{K_1+K_2} p_{vb}^{t_0,s,i} \approx 0 \end{array} \tag{17}$$

After t_0, since the new appearance of the background at pixel s is classified as foreground, $p_{b,v}^{t_k,s}$, $\sum_{i=1}^{K_1} p_v^{t_k,s,i}$ and $\sum_{i=1}^{K_1} p_{vb}^{t_k,s,i}$ decrease gradually, whereas $\sum_{i=K_1+1}^{K_1+K_2} p_v^{t_k,s,i}$ increase gradually and will be moved to the first K_2 positions in $S_{\mathbf{v}_t}^{s,t,i}$ by the re-sorting operation at each time step. Once the condition of (12) is met at time t_n, the new background state is learned. To make the expression simple and clear, let's suppose no re-sorting operation is performed at each time step. Then the condition (12) becomes

$$\sum_{i=K_1+1}^{K_1+K_2} p_v^{t_n,s,i} - p_{b,v}^{t_n,s}\sum_{i=K_1+1}^{K_1+K_2} p_{vb}^{t_n,s,i} > T \tag{18}$$

From (9) and (17), at time t_n after t_0 there are

$$p_{b,v}^{t_n,s} = (1-\alpha_2)^n p_{b,v}^{t_0,s} \approx (1-\alpha_2)^n \tag{19}$$

$$\begin{aligned} \sum_{i=K_1+1}^{K_1+K_2} p_v^{t_n,s,i} &= (1-\alpha_2)^n \sum_{i=K_1+1}^{K_1+K_2} p_v^{t_0,s,i} + \sum_{j=0}^{n-1}(1-\alpha_2)^j\alpha_2 \\ &\approx 1-(1-\alpha_2)^n \end{aligned} \tag{20}$$

$$\sum_{i=K_1+1}^{K_1+K_2} p_{vb}^{t_n,s,i} = (1-\alpha_2)^n \sum_{i=K_1+1}^{K_1+K_2} p_{vb}^{t_0,s,i} + 0 \approx 0 \tag{21}$$

By substituting the items in (18) with (19-21) and doing simple rearranging, we can obtain

$$\alpha_2 > 1 - (1-T)^{1/n} \tag{22}$$

This means, if we think that n frames for the system to response to an "once-off" background change is quick enough, we should choose the learning rate α_2 from (22). As example, if we want the system to response to an ideal "once-off" background change after 20 seconds with 25 fps frame rate and $T = 90\%$, α_2 should be larger than 0.0046 but not too larger than it to prevent the system not to sensitive to noise and foreground objects.

3.4.2 Updating the Reference Background Image

A reference background image that represents the most recent appearance of the background is also maintained at each time step to make the background difference accurate.

The gradual changes for stationary background objects are updated by using an Infinite Impulse Response (IIR) filter. If s is detected as a point of insignificant change in change detection, the reference background image is updated as

$$B_c(s, t+1) = (1 - \alpha_1)B_c(s, t) + \alpha_1 I_c(s, t) \qquad (23)$$

where $c \in \{r, g, b\}$ and α_1 is the parameter of the IIR filter. Since the next operation can adapt to "once-off" changes, a small positive number of α_1 is selected to smooth out the perturbs caused by image noise.

If $O(s, t) = 0$ and $F_{td}(s, t) = 1$ or $F_{bd}(s, t) = 1$, it indicates a background change is detected. The pixel is replaced by the new background appearance

$$B_c(s, t+1) = I_c(s, t), \text{ for } c = r, g, b \qquad (24)$$

With this operation, the reference background image can follow the background motion, e.g., the changes between tree branch and sky, and adapt to "once-off" changes in stationary background parts.

4. EXPERIMENTAL RESULTS

Experiments have been performed on many difficult videos of both indoor and outdoor scenes. The test videos include wavering tree branches and curtains, flickering screens, lights, or water surfaces, moving escalators, opening/closing doors, removed/deposited objects, switching on/off lights, lighting condition changes from day to night or clouds and raining, and shadows of people on the ground surface. Many of the test scenes are so difficult that less of the previous methods has been tested on. With the proposed method, the foreground objects were detected and segmented satisfactorily from these videos. Five examples from these difficult videos are presented in this paper. Meanwhile, quantitative evaluation of the performance and comparison with two existing methods has been performed. The first is a robust background subtraction (RBS) method [8] which is robust to various background variations and accurate to the camouflage of the foreground objects but a stationary background is required, and the second is the MoG method [11] which employs multiple color clusters for background modeling. For a fair comparison the same post-processing (i.e., the morphological smoothing and small region elimination) was used for the two methods.

In the experiments, our method was automatically initialized by setting all the prior and conditional probabilities to zero, i.e., $p_b^{s,0} = 0, p_v^{s,0,i} = 0, p_{vb}^{s,0,i} = 0$ for $i = 1, \cdots, N_2$ and $\mathbf{v}_t = \{\mathbf{c}_t, \mathbf{cc}_t\}$. The gradual updating process learns the most significant features gradually, and the "once-off" updating process converts the most significant features as background features if (12) is held. After the "once-off" learning, the system is able to work well for background and foreground classification.

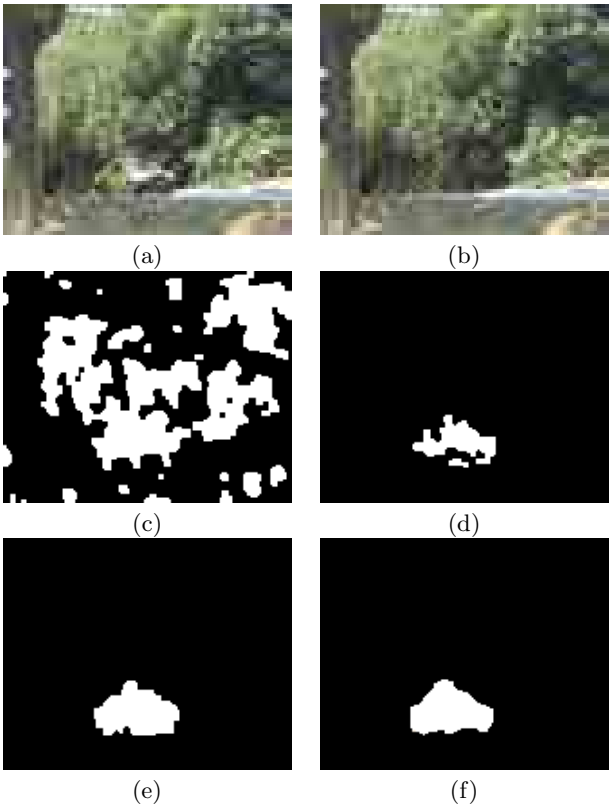

Figure 2: The test on a video containing wavering tree branches in strong winds. (a) a frame from the video, (b) the maintained background reference image, (c) the result of RBS, (d) the result of the MoG, (e) the result of the proposed method, (f) the "ground truth".

The segmentation results from five difficult videos are shown in Figure 2 to 6. In each figure, the displayed images are: (a) a frame from the video, (b) the background reference image maintained by the proposed method, (c) the result of RBS, (d) the result of MoG, (e) the result of the proposed method, and (f) the manually generated "ground truth".

The first example displayed in Figure 2 comes from a video containing moving tree bushes in strong wind. The great motion of tree branches can be seen from the result of RBS which is obtained by subtracting the frame from an empty frame just before the car moving into the scene. Comparing the segmentation result of the proposed method with the "ground truth", one can see that the foreground object was extracted accurately. As to MoG, since the coverage of background colors for moving trees are very large due to more than one cluster being used, many parts of the foreground object were mis-classified as the background.

The second example shown in Figure 3 displays human detection in front of the wavering curtains. The great motion of the curtains due to the wind can be seen from the result of RBS which is generated from the frame and an empty

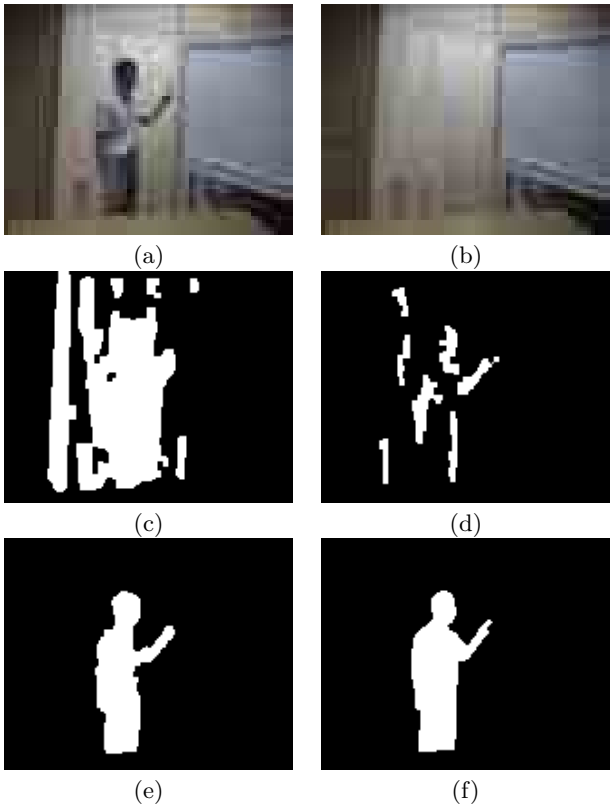

Figure 3: The test on a video containing moving curtains. (a) a frame from the video, (b) the maintained background reference image, (c) the result of RBS, (d) the result of the MoG, (e) the result of the proposed method, (f) the "ground truth".

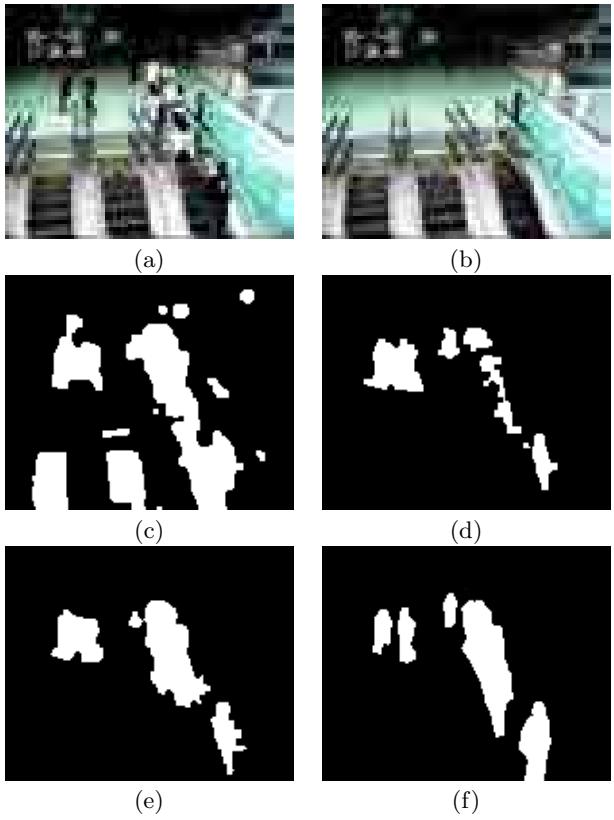

Figure 4: The test on a video from a subway station with moving escalators. (a) a frame from the video, (b) the maintained background reference image, (c) the result of RBS, (d) the result of the MoG, (e) the result of the proposed method, (f) the "ground truth".

frame just before the person coming into the scene. With the comparison to the "ground truth", one can see that the proposed method has accurately segmented the person from the video. RBS was too sensitive to moving background objects. The MoG was still sensitive to the motion of curtains. On the other hand, the color of curtains changes from dark to bright gray during the motion. When the two strips overlapped, the color is dark gray. Otherwise, the curtain is bright gray. Since the color of the trousers of the person is similar to the dark color of the curtains and the color of the shirt is similar to the bright color of the curtains, great part of the human body was mis-classified as background with MoG.

The third example in Figure 4 came from a video of subway station, which contains three moving escalators and human flow in the right sight of the images. The motion of the escalators can be observed in the result of RBS which is obtained from the frame and an empty frame with on persons in the scene from the video. Even though both MoG and the proposed method can suppress the moving escalators, the MoG turned to be less sensitive to the human objects on the escalators and the path of human flow since it learned too many colors for backgrounds in the motion regions of the images through the video. The proposed method segmented the human objects on both the moving escalator and ground surface quite well.

The fourth example which shows the segmentation of human objects in the presence of a large fountain is displayed in Figure 5. This video contains several persons walking in front of the fountain. With the color variation of the falling water, the RBS detected many false foreground pixels since the background is not stationary. On the other hand, MoG missed some foreground pixels with colors similar to the water. But the result of the proposed method was more accurate than the two existing methods because two types of the features were employed for the stationary and motion background parts, respectively.

The last example shown in Figure 6 tests the performance of the methods on a busy scene. In the example frame, there are four persons in the scene and the left man cast significant shadows on the ground surface. Compared with the "ground truth", the segmentation results can be evaluated visually as follows. For the results of RBS and the proposed method, four persons are detected well and the shadow of the left man is suppressed. As to the result of MoG, the right man and the great parts of the two middle persons were missed while the shadow of the left person was mis-detected as the foreground. In the maintained background reference image, the shadows of the persons can be observed while no person has been absorbed.

From these experimental results, it can be concluded as

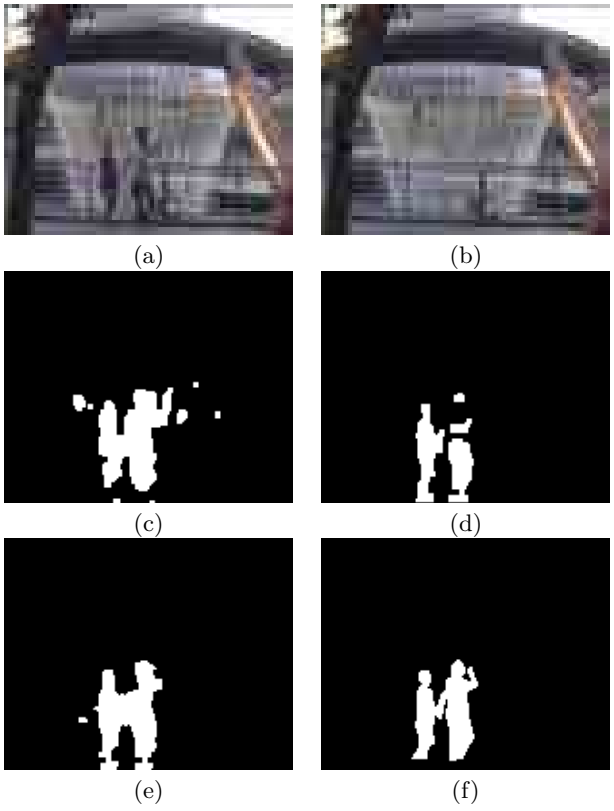

Figure 5: The test on a video containing a big fountain. (a) a frame from the video, (b) the maintained background reference image, (c) the result of RBS, (d) the result of the MoG, (e) the result of the proposed method, (f) the "ground truth".

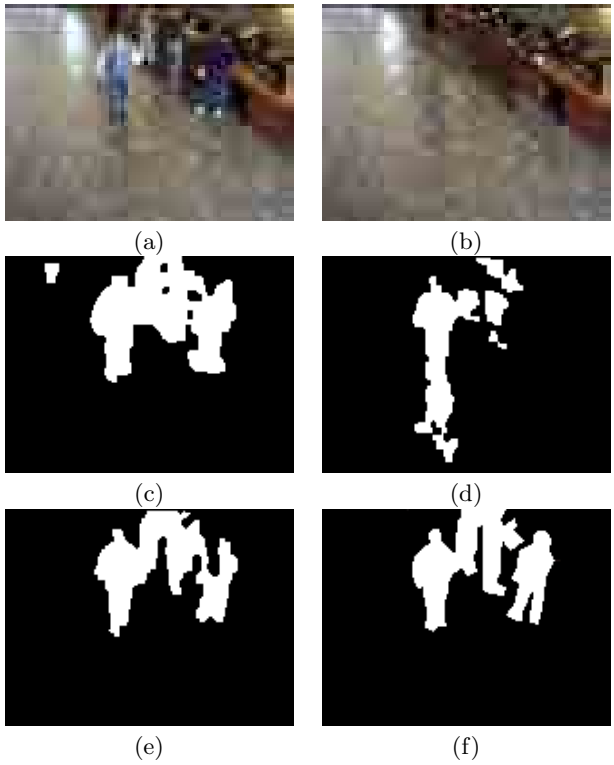

Figure 6: The test on a video of a busy scene. (a) a frame from the video, (b) the maintained background reference image, (c) the result of RBS, (d) the result of the MoG, (e) the result of the proposed method, (f) the "ground truth".

follows. The RBS method works well for a stationary background even there are variations of illumination changes, but it fails when the background involves motion objects. MoG can work for both stationary and motion background parts, but it often mis-classifies the foreground points as the background when there are moving background objects or frequent crowds in the video. The proposed method performs well for both stationary and motion background parts since it employs different significant features for different parts. These conclusions are more easy to be validated with the observations through the whole videos[1].

Quantitative evaluation of proposed method and comparison with two existing methods were also performed in this study. The results were evaluated quantitatively from the comparison with the "ground truths" in terms of:

- The False Negative Errors (F. Neg): the number of foreground points that are missed;
- The False Positive Errors (F. Pos): the number of background points that are mis-detected as foreground.

The quantitative evaluation results over the examples displayed in Figure 2 to 6 are shown in Table 1. The quantitative evaluation agrees with the conclusions from the visual observation of the experimental results.

There are a few parameters used in the proposed method, i.e., N_1, N_2, T, α_1, and α_2. However, since the decision of background and foreground classification using (4) is not directly dependent on any heuristic threshold, the performance of the proposed method is not sensitive to the parameters too much.

5. CONCLUSION

In this paper, we have proposed a novel method based on Bayes decision theory to detect foreground objects from complex videos which contain both stationary and moving background objects. A Bayes decision rule for background and foreground classification from a general feature vector is established. It is then applied to both stationary and moving background objects with suitable feature vectors. Different feature vectors are selected for different background parts. Foreground objects are segmented by fusing the results from both station and motion pixels. Both gradual and "once-off" learning strategies for learning and updating the feature statistics of background are proposed. The convergence of the learning algorithm is also analyzed. The proposed method has been tested on numerous real scenes containing wavering tree branches in strong winds, flickering screens/water surfaces, moving escalators, opening/closing doors, shadows of moving human objects, switching on/off

[1]Ten sequences (MPEG files) of test videos and results, including the examples presented in this paper, are available at http://perception.i2r.a-star.edu.sg

Table 1: Quantitative evaluation and comparison of the test results.

Sequence	Method	F. Neg	F. Pos	Total
Trees	Proposed	68	90	158
	MoG	370	58	428
	RBS	134	5814	5948
Curtain	Proposed	193	63	256
	MoG	1250	293	1543
	RBS	63	3533	3596
Escalator	Proposed	567	427	994
	MoG	1186	332	1518
	RBS	295	2451	2746
Fountain	Proposed	91	266	357
	MoG	324	117	441
	RBS	95	484	579
BusyScene	Proposed	401	268	669
	MoG	1311	615	1926
	RBS	161	786	947

lights, and illumination changes from day to night as well as clouds. Good results for foreground detection have been obtained from these scenes. The comparison with existing methods indicates that using different features to model different parts of background is more accurate than using just one type of features for complex background.

The weak point of the proposed method is that it is prone to absorb foreground objects if they are motionless for a long time. This is because the proposed method only learns the background features at pixel level. Further investigation is being conducted for improving the learning strategy by adding feedback control from high-level object recognition and tracking.

6. REFERENCES

[1] T. E. Boult, R. Micheals, X. Gao, P. Lewis, C. Power, W. Yin, and A. Erkan. Frame-rate omnidirectional surveillance & tracking of camouflaged and occluded targets. In *Proceedings IEEE Workshop on Visual Surveillance*, pages 48–55. IEEE Computer Society, 1999.

[2] X. Gao, T. Boult, F. Coetzee, and V. Ramesh. Error analysis of background adaption. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, pages 503–510. IEEE Computer Society, 2000.

[3] I. Haritaoglu, D. Harwood, and L. Davis. W^4: Real-time surveillance of people and their activities. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):809–830, August 2000.

[4] A. Iketani, A. Nagai, Y. Kuno, and Y. Shirai. Detecting persons on changing background. In *Proceedings of International Conference on Pattern Recognition*, pages 74–76, 1998.

[5] K. Karmann and A. V. Brandt. Moving object recognition using an adaptive background memory. *Time-Varing Image Processing and Moving Object Recognition*, 2:289–296, 1990.

[6] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russel. Toward robust automatic traffic scene analysis in real-time. In *Proceedings Int'l Conf. Pattern Recognition*, pages 126–131, 1994.

[7] L. Li, W. M. Huang, I. Y. H. Gu, and Q. Tian. Foreground object detection in changing background based on color co-occurrenece statistics. In *Proceedings IEEE Workshop on Application of Computer Vision*, pages 269–274, 2002.

[8] L. Li and M. K. H. Leung. Integrating intensity and texture differences for robust change detection. *IEEE Trans. Image Processing*, 11(2):105–112, February 2002.

[9] A. Lipton, H. Fujiyoshi, and R. Patil. Moving target classification and tracking from real-time video. In *Proceedings IEEE Workshop on Application of Computer Vision*, pages 8–14. IEEE Computer Society, 1998.

[10] P. Rosin. Thresholding for change detection. In *Proceedings of IEEE Int'l Conf. on Computer Vision*, pages 274–279, 1998.

[11] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):747–757, August 2000.

[12] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Proceedings of IEEE Int'l Conf. on Computer Vision*, pages 255–261. IEEE Computer Society, 1999.

[13] L. Wixson. Detecting salient motion by accumulating directionary-consistenct flow. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):774–780, August 2000.

[14] C. Wren, A. Azarbaygaui, T. Darrell, and A. Pentland. *Pfinder*: Real-time tracking of the human body. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997.