

A probabilistic approach for foreground and shadow segmentation in monocular image sequences

Yang Wang^{a, b, *}, Tele Tan^c, Kia-Fock Loe^b, Jian-Kang Wu^a

^a*Institute for Infocomm Research, Singapore 119613*

^b*Department of Computer Science, National University of Singapore, Singapore 117543*

^c*Department of Computing, Curtin University of Technology, Western Australia 6102, Australia*

Received 2 October 2003; received in revised form 17 February 2005; accepted 17 February 2005

Abstract

This paper presents a novel method of foreground and shadow segmentation in monocular indoor image sequences. The models of background, edge information, and shadow are set up and adaptively updated. A Bayesian network is proposed to describe the relationships among the segmentation label, background, intensity, and edge information. A maximum a posteriori—Markov random field estimation is used to boost the spatial connectivity of segmented regions.

© 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Bayesian network; Foreground segmentation; Graphical model; Markov random field; Shadow detection

1. Introduction

Detecting dynamic objects in image sequences is very important in application areas such as surveillance and object-based coding. Effective and efficient background removal is critical in these systems. Background subtraction based on intensity or color is a commonly used technique to detect foreground objects. The background model is built from observed images and foreground elements are identified if they show significant difference from the background.

To deal with illumination or object changes in the background, many researchers [1,2] have abandoned nonadaptive methods of backgrounding. The accumulation of errors in the background over time makes the method useful only in stationary environments. Friedman and Russell [3]

classify each pixel by a probabilistic model of how that pixel looks when it is part of different classes and use an incremental EM algorithm to learn the pixel model. Stauffer and Grimson [4] model each pixel as a mixture of Gaussians and update the model in an adaptive way. The Gaussian distributions are then evaluated to determine which are possibly from a background process. Elgammal et al. [5] employ kernel density estimation for nonparametric background modeling. Recently, hidden Markov models [6,7] have been used to model the dynamical dependencies in the background process.

Besides the nonstationariness of the background, camouflage and shadow are two classic problems of subtraction. If foreground objects have similar colors as the background, they may be erroneously removed from the scene. In addition, moving shadows cast on the background may be erroneously detected as foreground [8]. Depth computation from stereo cameras can be used to handle these two problems [9]. For monocular color video sequences, false segmentation caused by shadows can be reduced by computing differences in a normalized color space that is insensitive to illumination change [10,11]. Moreover, edge information

* Corresponding author. Institute for Infocomm Research, Singapore 119613. Tel.: +65 9478 7256.

E-mail addresses: yang.wang@ieee.org (Y. Wang), teletan@cs.curtin.edu.au (T. Tan), loekf@comp.nus.edu.sg (K.-F. Loe), jiankang@i2r.a-star.edu.sg (J.-K. Wu).

can be employed to improve the reliability of the results [12]. Stauder et al. [13] assume that static edges in the background remain under shadow and penumbras exist at the boundary of shadows. However, this is sometimes not true due to the properties of the imaging process. Mikic et al. [14] instead approximate the change of the camera response for the shadowed region by a diagonal matrix.

On the other hand, graphical probabilistic models provide a natural tool for handling uncertainty and complexity through the combination of probability theory and graph theory. In particular, Bayesian belief networks and Markov random fields are playing increasingly important roles in the design and analysis of machine intelligent systems [15]. Graphical models have attracted more and more attention in vision applications such as traffic scene analysis [16,17], layer extraction from image sequences [18], and human motion tracking [19].

To solve the above mentioned problems in monocular indoor grayscale sequences, a unified framework of foreground segmentation is proposed in this paper. We introduce a Bayesian network to combine the background, intensity, and edge information. A generalized model is built for the appearance change under shadow. Camouflage is decreased by encouraging the formation of continuous segmentation regions. Parameters in the model can be updated adaptively. The solution is obtained by maximizing the posterior probability density of the segmentation field using a noniterative algorithm. Experiments show that our method greatly improves the accuracy of segmentation. The rest part of the paper is arranged as follows: Section 2 presents the formulation of the models. Sections 3 and 4 describes the segmentation method. Section 5 proposes the implementation details. Section 6 discusses the experimental results. At the end, our technique is concluded in Section 7.

2. Model representation

Given the image sequence $\{g_k\}$, we would like to classify each pixel of each image as foreground (moving object), shadow, or background. The segmentation label for a point is defined as

$$s_k(\mathbf{x}) = \begin{cases} 1, & \text{if site } \mathbf{x} \text{ is in the background,} \\ 2, & \text{if site } \mathbf{x} \text{ is shadowed by the foreground,} \\ 3, & \text{if site } \mathbf{x} \text{ is in the foreground,} \end{cases}$$

$$\forall \mathbf{x} \in \mathbf{X}, k = 1, 2, \dots,$$

where $s_k(\mathbf{x})$ is the label of a single pixel \mathbf{x} within the image at time k , and \mathbf{X} is the spatial domain of the video scene. Static shadows are considered to be part of the background. The entire segmentation field is expressed compactly as s_k .

2.1. Background model

In order to segment the foreground in a video sequence, the system must first model the background of the video

scene. Each image acquired by the camera contains noise components. Assume that independent Gaussian noise corrupts each pixel in the scene, so that the observation model for the background becomes

$$b_k(\mathbf{x}) = \mu_{b,k}(\mathbf{x}) + n_k(\mathbf{x}), \quad (1)$$

where random variable $b_k(\mathbf{x})$ is the intensity of a single pixel \mathbf{x} within the background at time k , and $\mu_{b,k}(\mathbf{x})$ is the intensity mean. $n_k(\mathbf{x})$ is the independent zero-mean additive noise with variance $\sigma_{b,k}^2(\mathbf{x})$ at time k . The parameter vector $(\mu_{b,k}(\mathbf{x}), \sigma_{b,k}^2(\mathbf{x}))^T$ is denoted as $\boldsymbol{\theta}_{b,k}(\mathbf{x})$, and the entire background is expressed as $\boldsymbol{\theta}_{b,k}$. For each site \mathbf{x} in the background, the mean intensity and variance at time k could be estimated from its history. In this work, the intensity of each point is represented by its grayscale value, which ranges from 0 to $y_{\max} = 255$.

2.2. Edge model

The edge model is built by applying the edge operator to the scene, which produces a horizontal difference image and a vertical difference image. For the k th frame g_k , $\mathbf{e}_{g,k}(\mathbf{x})$ is the edge vector at site $\mathbf{x} = (x_1, x_2)$,

$$\mathbf{e}_{g,k}(\mathbf{x}) = (e_{g,k}^h(\mathbf{x}), e_{g,k}^v(\mathbf{x}))^T, \quad (2a)$$

$$e_{g,k}^h(\mathbf{x}) = g_k(x_1 + 1, x_2) - g_k(x_1 - 1, x_2), \quad (2b)$$

$$e_{g,k}^v(\mathbf{x}) = g_k(x_1, x_2 + 1) - g_k(x_1, x_2 - 1), \quad (2c)$$

where $g_k(\mathbf{x})$ is the intensity of a single point \mathbf{x} within the k th video frame, $e_{g,k}^h(\mathbf{x})$ and $e_{g,k}^v(\mathbf{x})$ are the horizontal difference and vertical difference, respectively. The entire difference image is expressed as $\mathbf{e}_{g,k}$.

Similarly, we can define the edge information for the background,

$$\mathbf{e}_{b,k}(\mathbf{x}) = (e_{b,k}^h(\mathbf{x}), e_{b,k}^v(\mathbf{x}))^T, \quad (3a)$$

$$e_{b,k}^h(\mathbf{x}) = b_k(x_1 + 1, x_2) - b_k(x_1 - 1, x_2), \quad (3b)$$

$$e_{b,k}^v(\mathbf{x}) = b_k(x_1, x_2 + 1) - b_k(x_1, x_2 - 1). \quad (3c)$$

From the background model we know that $\mathbf{e}_{b,k}(\mathbf{x})$ is of bivariate normal distribution with mean difference $\boldsymbol{\mu}_{\mathbf{e},k}(\mathbf{x})$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{e},k}(\mathbf{x})$ for each site \mathbf{x} . $\boldsymbol{\mu}_{\mathbf{e},k}(\mathbf{x})$ is determined by the intensity means of the four neighboring points,

$$E[e_{b,k}^h(\mathbf{x})] = \mu_{b,k}(x_1 + 1, x_2) - \mu_{b,k}(x_1 - 1, x_2), \quad (4a)$$

$$E[e_{b,k}^v(\mathbf{x})] = \mu_{b,k}(x_1, x_2 + 1) - \mu_{b,k}(x_1, x_2 - 1). \quad (4b)$$

By the independent noise assumption in the background model, $\boldsymbol{\Sigma}_{\mathbf{e},k}(\mathbf{x})$ can be calculated from the variances of the neighboring points,

$$\text{Var}[e_{b,k}^h(\mathbf{x})] = \sigma_{b,k}^2(x_1 + 1, x_2) + \sigma_{b,k}^2(x_1 - 1, x_2), \quad (5a)$$

$$\text{Var}[e_{b,k}^v(\mathbf{x})] = \sigma_{b,k}^2(x_1, x_2 + 1) + \sigma_{b,k}^2(x_1, x_2 - 1), \quad (5b)$$

$$\text{Cov}[e_{b,k}^h(\mathbf{x}), e_{b,k}^v(\mathbf{x})] = 0. \quad (5c)$$

The parameter vector $(\boldsymbol{\mu}_{e,k}(\mathbf{x}), \boldsymbol{\Sigma}_{e,k}(\mathbf{x}))^T$ is denoted as $\boldsymbol{\theta}_{e,k}(\mathbf{x})$, and the entire field at time k is expressed as $\boldsymbol{\theta}_{e,k}$. The edge model can be used to locate changes in the structure of the scenes as edges appear, vanish, or rotate.

2.3. Shadow model

Given the background intensity of a point \mathbf{x} when illuminated, we use a linear transformation to approximate the change of intensity for the same point when shadowed in the video frame at time k ,

$$g_k(\mathbf{x}) = a_k b_k(\mathbf{x}) + c_k, \quad \text{if } s_k(\mathbf{x}) = 2. \quad (6)$$

When a_k equals 1, the edge information will not change if the area is shadowed by the foreground. Moreover, if we extend the image input from one-channel (grayscale) to multi-channel (R, G, B), the chromaticity [10,17] will remain unchanged under such a linear transformation when c_k is zero. So the shadow model can be viewed as the generalization of the previous assumptions. With this model for the appearance change, we can easily estimate means and variances for the points under shadow. For a point \mathbf{x} under shadow, its intensity mean is $a_k \mu_{b,k}(\mathbf{x}) + c_k$, and its variance is $a_k^2 \sigma_{b,k}^2(\mathbf{x})$ at time k . Thus the mean of pixel intensity under shadow is controlled by a_k and c_k , and the variance is controlled by a_k . At the beginning a_0 and c_0 are manually initialized according to the visual environment, then parameters a_k and c_k are adaptively updated over time (see Section 5.1).

3. Adaptive backgrounding

For static background, a sequence of background images may be recorded at the beginning and the intensity mean and variance of each pixel can be calculated.

For nonstationary background, the update method is based on the ideas from Stauffer et al. [4] and Harville et al. [20]. The recent history of each pixel, $\{g_i(\mathbf{x})\}_{1 \leq i \leq k}$, is modeled by a mixture of Gaussian distributions. The probability of the current observation is

$$p(g_k(\mathbf{x})) = \sum_{i=1}^K w_{i,k}(\mathbf{x}) p(g_k(\mathbf{x}) | \mu_{i,k}(\mathbf{x}), \sigma_{i,k}^2(\mathbf{x})), \quad (7a)$$

$$\begin{aligned} p(g_k(\mathbf{x}) | \mu_{i,k}(\mathbf{x}), \sigma_{i,k}^2(\mathbf{x})) \\ = \frac{1}{\sqrt{2\pi}\sigma_{i,k}(\mathbf{x})} \exp \left\{ -\frac{1}{2\sigma_{i,k}^2(\mathbf{x})} [g_k(\mathbf{x}) - \mu_{i,k}(\mathbf{x})]^2 \right\}, \end{aligned} \quad (7b)$$

where K is the number of distributions (usually from three to five are used), $w_{i,k}(\mathbf{x})$ is the normalized weight of the i th

Gaussian in the mixture at time k , $\mu_{i,k}(\mathbf{x})$ and $\sigma_{i,k}^2(\mathbf{x})$ are the mean and variance of the i th Gaussian at time k .

At current time k , each new value $g_k(\mathbf{x})$ is checked to match the existing Gaussian distributions (the value is matched if it is within 3 standard deviations of a distribution). If the i th Gaussian is found to match the new value, its distribution parameters are updated as follows:

$$w_{i,k}(\mathbf{x}) = (1 - \alpha)w_{i,k-1}(\mathbf{x}) + \alpha, \quad (8a)$$

$$\mu_{i,k}(\mathbf{x}) = (1 - \alpha)\mu_{i,k-1}(\mathbf{x}) + \alpha g_k(\mathbf{x}), \quad (8b)$$

$$\sigma_{i,k}^2(\mathbf{x}) = (1 - \alpha)\sigma_{i,k-1}^2(\mathbf{x}) + \alpha(g_k(\mathbf{x}) - \mu_{i,k-1}(\mathbf{x}))^2, \quad (8c)$$

where α is the learning rate. Eq. (8) is equivalent to the expectation with an exponential factor for the past values. For unmatched distributions, the means and variances remain the same, while the weights should be renormalized. If none of the distributions are matched, the distribution of the lowest weight is replaced with a Gaussian with the new value as its mean, initially low weight and high variance.

As the parameters of the mixture model change, the Gaussian distribution that has the highest ratio of weight over variance is chosen as the background model for each site.

$$\boldsymbol{\theta}_{b,k}(\mathbf{x}) = (\mu_{m_{\mathbf{x}},k}(\mathbf{x}), \sigma_{m_{\mathbf{x}},k}^2(\mathbf{x}))^T, \quad (9)$$

where $m_{\mathbf{x}} = \arg \max_i w_{i,k}(\mathbf{x}) / \sigma_{i,k}(\mathbf{x})$. Each time after background updating, the background edge information $\boldsymbol{\theta}_{e,k}$ at time k can be calculated by Eqs. (4) and (5).

4. Bayesian foreground detection

To extract the foreground given the current frame g_k , difference image $\mathbf{e}_{g,k}$, background $\boldsymbol{\theta}_{b,k}$, and background edge information $\boldsymbol{\theta}_{e,k}$, we wish to compute the maximum a posteriori (MAP) estimation of the segmentation field s_k . Using the Bayes' rule and ignoring the constants with respect to the unknowns,

$$\begin{aligned} \hat{s}_k &= \arg \max_{s_k} p(s_k | \boldsymbol{\theta}_{b,k}, \boldsymbol{\theta}_{e,k}, g_k, \mathbf{e}_{g,k}) \\ &= \arg \max_{s_k} p(s_k, \boldsymbol{\theta}_{b,k}, \boldsymbol{\theta}_{e,k}, g_k, \mathbf{e}_{g,k}) \\ &= \arg \max_{s_k} p(\boldsymbol{\theta}_{b,k}, \boldsymbol{\theta}_{e,k}, g_k, \mathbf{e}_{g,k} | s_k) p(s_k), \end{aligned} \quad (10)$$

where $\boldsymbol{\theta}_{b,k}$ is defined in Section 2.1, $\mathbf{e}_{g,k}$ and $\boldsymbol{\theta}_{e,k}$ are described in Section 2.2. The likelihood model $p(\boldsymbol{\theta}_{b,k}, \boldsymbol{\theta}_{e,k}, g_k, \mathbf{e}_{g,k} | s_k)$ and the prior model $p(s_k)$ must be defined for the video sequence.

4.1. Likelihood model

Assuming conditional independence among spatially distinct observations, we factorize the likelihood model as

$$\begin{aligned} p(\boldsymbol{\theta}_{b,k}, \boldsymbol{\theta}_{e,k}, g_k, \mathbf{e}_{g,k} | s_k) \\ = \prod_{\mathbf{x} \in \mathbf{X}} p(\boldsymbol{\theta}_{b,k}(\mathbf{x}), \boldsymbol{\theta}_{e,k}(\mathbf{x}), g_k(\mathbf{x}), \mathbf{e}_{g,k}(\mathbf{x}) | s_k(\mathbf{x})). \end{aligned} \quad (11)$$

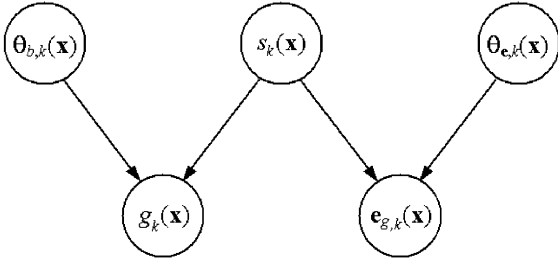


Fig. 1. A Bayesian network for foreground segmentation.

The relationships among $s_k(\mathbf{x})$, $\theta_{b,k}(\mathbf{x})$, $\theta_{e,k}(\mathbf{x})$, $g_k(\mathbf{x})$, and $e_{g,k}(\mathbf{x})$ can be modeled by a Bayesian network in Fig. 1. Given the segmentation label, background, and background edge information at the site, we assume that the image intensity is independent on the image edge. The conditional independence relationships implied by the belief network allow us to represent the joint more compactly [21]. Using the chain rule, the likelihood can be factorized as the product of the intensity likelihood $p(g_k(\mathbf{x})|\theta_{b,k}(\mathbf{x}), s_k(\mathbf{x}))$ and edge likelihood $p(e_{g,k}(\mathbf{x})|\theta_{e,k}(\mathbf{x}), s_k(\mathbf{x}))$ at site \mathbf{x} .

$$\begin{aligned} p(\theta_{b,k}(\mathbf{x}), \theta_{e,k}(\mathbf{x}), g_k(\mathbf{x}), e_{g,k}(\mathbf{x})|s_k(\mathbf{x})) \\ = p(\theta_{b,k}(\mathbf{x}))p(\theta_{e,k}(\mathbf{x}))p(g_k(\mathbf{x})|\theta_{b,k}(\mathbf{x}), s_k(\mathbf{x})) \\ \times p(e_{g,k}(\mathbf{x})|\theta_{e,k}(\mathbf{x}), s_k(\mathbf{x})) \\ \propto p(g_k(\mathbf{x})|\theta_{b,k}(\mathbf{x}), s_k(\mathbf{x}))p(e_{g,k}(\mathbf{x})|\theta_{e,k}(\mathbf{x}), s_k(\mathbf{x})). \end{aligned} \quad (12)$$

When site \mathbf{x} is labeled as the background, we can calculate the intensity likelihood model $p(g_k(\mathbf{x})|\theta_{b,k}(\mathbf{x}), s_k(\mathbf{x}))$ using the background model,

$$\begin{aligned} p(g_k(\mathbf{x})|\theta_{b,k}(\mathbf{x}), s_k(\mathbf{x}) = 1) \\ = \frac{1}{\sqrt{2\pi}\sigma_{b,k}(\mathbf{x})} \exp \left\{ -\frac{1}{2\sigma_{b,k}^2(\mathbf{x})} [g_k(\mathbf{x}) - \mu_{b,k}(\mathbf{x})]^2 \right\}. \end{aligned} \quad (13)$$

When site \mathbf{x} is shadowed, the density can be calculated by the shadow model,

$$\begin{aligned} p(g_k(\mathbf{x})|\theta_{b,k}(\mathbf{x}), s_k(\mathbf{x}) = 2) \\ = \frac{1}{\sqrt{2\pi}a_k\sigma_{b,k}(\mathbf{x})} \\ \times \exp \left\{ -\frac{1}{2a_k^2\sigma_{b,k}^2(\mathbf{x})} [g_k(\mathbf{x}) - a_k\mu_{b,k}(\mathbf{x}) - c_k]^2 \right\}. \end{aligned} \quad (14)$$

When site \mathbf{x} is labeled as the foreground, the background has no contribution to the image intensity information. Uniform distribution is assumed for the pixel. The conditional probability density becomes

$$\begin{aligned} p(g_k(\mathbf{x})|\theta_{b,k}(\mathbf{x}), s_k(\mathbf{x}) = 3) = p(g_k(\mathbf{x})|s_k(\mathbf{x}) = 3) \\ = \frac{1}{y_{\max}}. \end{aligned} \quad (15)$$

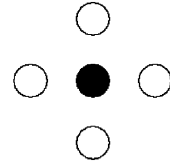


Fig. 2. The first-order neighborhood system.

Here $[0, y_{\max}]$ ($y_{\max} = 255$) is the range of grayscale value for pixel intensity.

For each point \mathbf{x} , denote the set of its four nearest neighboring points by $M_{\mathbf{x}}$ (the first-order neighborhood, see Fig. 2). Consider the spatial connectivity of the image, we assume the neighboring points have the same segmentation labels. Thus the edge likelihood $p(e_{g,k}(\mathbf{x})|\theta_{e,k}(\mathbf{x}), s_k(\mathbf{x}))$ can be approximated by

$$\begin{aligned} p(e_{g,k}(\mathbf{x})|\theta_{e,k}(\mathbf{x}), s_k(\mathbf{x})) \\ \approx p(e_{g,k}(\mathbf{x})|\theta_{e,k}(\mathbf{x}), s_k(\mathbf{y}) = s_k(\mathbf{x}), \forall \mathbf{y} \in M_{\mathbf{x}}) \\ = p \left(e_{g,k}(\mathbf{x})|\theta_{e,k}(\mathbf{x}), \prod_{\mathbf{y} \in M_{\mathbf{x}}} s_k(\mathbf{y}) = s_k(\mathbf{x})^{|M_{\mathbf{x}}|} \right), \end{aligned} \quad (16)$$

where $|M_{\mathbf{x}}|$ is the number of elements in the set.

Similarly, when the neighborhood area $M_{\mathbf{x}}$ belongs to the background, the density can be computed by the edge model,

$$\begin{aligned} p \left(e_{g,k}(\mathbf{x})|\theta_{e,k}(\mathbf{x}), \prod_{\mathbf{y} \in M_{\mathbf{x}}} s_k(\mathbf{y}) = 1 \right) \\ = \frac{1}{2\pi\sqrt{|\Sigma_{e,k}(\mathbf{x})|}} \exp \left\{ -\frac{1}{2} [e_{g,k}(\mathbf{x}) - \mu_{e,k}(\mathbf{x})]^T \right. \\ \left. \times \Sigma_{e,k}^{-1}(\mathbf{x}) [e_{g,k}(\mathbf{x}) - \mu_{e,k}(\mathbf{x})] \right\}. \end{aligned} \quad (17)$$

When the neighborhood area $M_{\mathbf{x}}$ is shadowed, the density can be computed from the shadow model,

$$\begin{aligned} p \left(e_{g,k}(\mathbf{x})|\theta_{e,k}(\mathbf{x}), \prod_{\mathbf{y} \in M_{\mathbf{x}}} s_k(\mathbf{y}) = 2^{|M_{\mathbf{x}}|} \right) \\ = \frac{1}{2\pi\sqrt{|a_k^2\Sigma_{e,k}(\mathbf{x})|}} \exp \left\{ -\frac{1}{2a_k^2} [e_{g,k}(\mathbf{x}) - a_k\mu_{e,k}(\mathbf{x})]^T \right. \\ \left. \times \Sigma_{e,k}^{-1}(\mathbf{x}) [e_{g,k}(\mathbf{x}) - a_k\mu_{e,k}(\mathbf{x})] \right\}. \end{aligned} \quad (18)$$

When neighborhood area $M_{\mathbf{x}}$ belongs to the foreground, we assume that the points within $M_{\mathbf{x}}$ are independent and

identically distributed (i.i.d.). From (15), we know

$$\begin{aligned}
 & p \left(\mathbf{e}_{g,k}(\mathbf{x}) | \boldsymbol{\theta}_{\mathbf{e},k}(\mathbf{x}), \prod_{\mathbf{y} \in M_{\mathbf{x}}} s_k(\mathbf{y}) = 3^{|\mathbf{M}_{\mathbf{x}}|} \right) \\
 &= p \left(\mathbf{e}_{g,k}(\mathbf{x}) \middle| \prod_{\mathbf{y} \in M_{\mathbf{x}}} s_k(\mathbf{y}) = 3^{|\mathbf{M}_{\mathbf{x}}|} \right) \\
 &= p \left(\mathbf{e}_{g,k}^h(\mathbf{x}) \middle| \prod_{\mathbf{y} \in M_{\mathbf{x}}} s_k(\mathbf{y}) = 3^{|\mathbf{M}_{\mathbf{x}}|} \right) \\
 &\quad \times p \left(\mathbf{e}_{g,k}^v(\mathbf{x}) \middle| \prod_{\mathbf{y} \in M_{\mathbf{x}}} s_k(\mathbf{y}) = 3^{|\mathbf{M}_{\mathbf{x}}|} \right) \\
 &= \left(\frac{1}{y_{\max}} - \frac{|\mathbf{e}_{g,k}^h(\mathbf{x})|}{y_{\max}^2} \right) \left(\frac{1}{y_{\max}} - \frac{|\mathbf{e}_{g,k}^v(\mathbf{x})|}{y_{\max}^2} \right). \quad (19)
 \end{aligned}$$

4.2. Prior model

The prior model $p(s_k)$ represents the prior probability of the segmentation field. We model the density by a Markov random field [22]. That is, if $N_{\mathbf{x}}$ is the neighborhood of a pixel \mathbf{x} , then the conditional distribution of a single label at \mathbf{x} completely depends on the labels within its neighborhood $N_{\mathbf{x}}$. According to the Hammersley–Clifford theorem, the density is given by a Gibbs distribution with the following form [23]:

$$p(s_k) \propto \exp \left\{ - \sum_{c \in C} V_k(s_k(\mathbf{x}) | \mathbf{x} \in c) \right\}, \quad (20)$$

where C is the set of all cliques c , and V_k is the clique potential function at time k . A clique is a set of pixels that are neighbors of each other. The clique potential depends only on the pixels within clique c . Only one-pixel and two-pixel cliques are used in our work.

The single-pixel clique potentials can be defined as

$$V_{1,k}(s_k(\mathbf{x})) = \eta_{s_k(\mathbf{x}),k}. \quad (21)$$

They reflect our prior knowledge of the probabilities of different region types. The lower the value of $\eta_{s_k(\mathbf{x}),k}$, the more likely that a point \mathbf{x} is labeled as $s_k(\mathbf{x})$ at time k .

Spatial connectivity can be imposed by the following two-pixel clique potential:

$$V_2(s_k(\mathbf{x}), s_k(\mathbf{y})) = \frac{1}{\|\mathbf{x} - \mathbf{y}\|^2} (1 - \delta(s_k(\mathbf{x}) - s_k(\mathbf{y}))), \quad (22)$$

where $\delta(\cdot)$ is the Kronecker delta function, and $\|\cdot\|$ denotes the Euclidian distance. Thus two neighboring pixels are more likely to belong to the same class than to different classes. The constraint becomes stronger with decrease of the distance between the neighboring sites.

Combining the above models, the Bayesian MAP estimate is obtained by minimizing the objective function

$$\begin{aligned}
 F_k(s_k) &= \sum_{\mathbf{x} \in \mathbf{X}} U_{1,k}(\mathbf{x}, s_k(\mathbf{x})) + \sum_{\mathbf{x} \in \mathbf{X}} U_{2,k}(\mathbf{x}, s_k(\mathbf{x})) \\
 &\quad + \lambda_1 \sum_{\mathbf{x} \in \mathbf{X}} V_{1,k}(s_k(\mathbf{x})) \\
 &\quad + \lambda_2 \sum_{\{\mathbf{x}, \mathbf{y}\} \in C} V_2(s_k(\mathbf{x}), s_k(\mathbf{y})), \quad (23)
 \end{aligned}$$

where $U_{1,k}(\mathbf{x}, s_k(\mathbf{x})) = -\ln p(g_k(\mathbf{x}) | \boldsymbol{\theta}_{b,k}(\mathbf{x}), s_k(\mathbf{x}))$, and $U_{2,k}(\mathbf{x}, s_k(\mathbf{x})) = -\ln p(\mathbf{e}_{g,k}(\mathbf{x}) | \boldsymbol{\theta}_{\mathbf{e},k}(\mathbf{x}), s_k(\mathbf{x}))$. $p(g_k(\mathbf{x}) | \boldsymbol{\theta}_{b,k}(\mathbf{x}), s_k(\mathbf{x}))$ and $p(\mathbf{e}_{g,k}(\mathbf{x}) | \boldsymbol{\theta}_{\mathbf{e},k}(\mathbf{x}), s_k(\mathbf{x}))$ are the likelihood models for intensity and edge respectively. The parameters $\eta_{1,k}, \eta_{2,k}, \eta_{3,k}, \lambda_1$ and λ_2 should be determined carefully to control the influence of each term in (23).

5. Implementation

5.1. Parameter determination

After the segmentation of the k th frame, denote the set of points labeled as s ($s = 1, 2, 3$) by $\mathbf{X}_{s,k}$. The single-pixel clique potential can be reestimated as

$$\eta_{i,k}^* = - \frac{|\mathbf{X}_{i,k}|}{\sum_s |\mathbf{X}_{s,k}|}, \quad i = 1, 2, 3. \quad (24)$$

With the learning rate α , $\eta_{i,k+1}$ can be updated in an adaptive way.

$$\eta_{i,k+1} = (1 - \alpha)\eta_{i,k} + \alpha\eta_{i,k}^*. \quad (25)$$

The parameters of the linear transformation in the shadow model can be reestimated from the set $\mathbf{X}_{2,k}$ by the least squares method,

$$a_k^* = \frac{\sum_{\mathbf{x} \in \mathbf{X}_{2,k}} g_k(\mathbf{x}) \sum_{\mathbf{x} \in \mathbf{X}_{2,k}} \mu_{b,k}(\mathbf{x}) - |\mathbf{X}_{2,k}| \sum_{\mathbf{x} \in \mathbf{X}_{2,k}} g_k(\mathbf{x}) \mu_{b,k}(\mathbf{x})}{(\sum_{\mathbf{x} \in \mathbf{X}_{2,k}} \mu_{b,k}(\mathbf{x}))^2 - |\mathbf{X}_{2,k}| \sum_{\mathbf{x} \in \mathbf{X}_{2,k}} \mu_{b,k}^2(\mathbf{x})}, \quad (26a)$$

$$c_k^* = \frac{\sum_{\mathbf{x} \in \mathbf{X}_{2,k}} g_k(\mathbf{x}) - a_k^* \sum_{\mathbf{x} \in \mathbf{X}_{2,k}} \mu_{b,k}(\mathbf{x})}{|\mathbf{X}_{2,k}|}. \quad (26b)$$

The shadow model is then updated adaptively.

$$a_{k+1} = (1 + \eta_{2,k}^* \alpha) a_k - \eta_{2,k}^* \alpha a_k^*, \quad (27a)$$

$$c_{k+1} = (1 + \eta_{2,k}^* \alpha) c_k - \eta_{2,k}^* \alpha c_k^*. \quad (27b)$$

In (27) the effective learning rate $-\eta_{2,k}^* \alpha$ changes with the ratio of shadowed points in the scene. This helps to make a robust updating process, especially for the frames where there are only few shadowed points.

Parameters α , λ_1 , and λ_2 are manually determined to reflect the importance of previous knowledge, one-pixel clique potential, and two-pixel clique potential respectively.

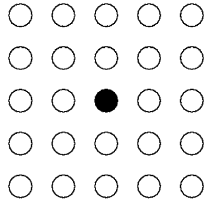


Fig. 3. The fifth-order neighborhood system.

5.2. Optimization

Obviously, there is no simple method of performing the optimization in (23), furthermore, the objective function does not have a unique minimum since it is nonconvex in terms of $s_k(\mathbf{x})$. To arrive at a sub-optimal estimate, we use a local technique known as highest confidence first (HCF). HCF is a noniterative and deterministic algorithm that guarantees to reach a local optimum after a finite number of steps [24]. Its feature is the introduction of a special uncommitted label 0 in the labeling strategy, so that the original label set is augmented by this label into $\{0, 1, 2, 3\}$.

Given the labels of the points within the neighborhood $N_{\mathbf{x}}$, the conditional posterior potential for a point \mathbf{x} at time k is defined as

$$f_k(\mathbf{x}, s_k(\mathbf{x})) = U_{1,k}(\mathbf{x}, s_k(\mathbf{x})) + U_{2,k}(\mathbf{x}, s_k(\mathbf{x})) + \lambda_1 V_{1,k}(s_k(\mathbf{x})) + \lambda_2 \sum_{\mathbf{y} \in N_{\mathbf{x}}} V_2^*(s_k(\mathbf{x}), s_k(\mathbf{y})), \quad (28a)$$

$$V_2^*(s_k(\mathbf{x}), s_k(\mathbf{y})) = \begin{cases} 0 & \text{if } s_k(\mathbf{y})=0, \\ V_2(s_k(\mathbf{x}), s_k(\mathbf{y})) & \text{otherwise.} \end{cases} \quad (28b)$$

In our work, the fifth-order neighborhood system is used (see Fig. 3). Based on the conditional posterior potential, we can define the stability measure of site \mathbf{x} .

$$S_k(\mathbf{x}, s_k(\mathbf{x})) = \begin{cases} -\min_{s \neq 0, s_{\min,k}(\mathbf{x})} [f_k(\mathbf{x}, s) - f_k(\mathbf{x}, s_{\min,k}(\mathbf{x}))] & \text{if } s_k(\mathbf{x})=0, \\ \min_{s \neq 0, s_k(\mathbf{x})} [f_k(\mathbf{x}, s) - f_k(\mathbf{x}, s_k(\mathbf{x}))] & \text{otherwise,} \end{cases} \quad (29)$$

where $s_{\min,k}(\mathbf{x}) = \arg \min_{s \neq 0} f_k(\mathbf{x}, s)$.

The stability measure [25], i.e. $S_k(\mathbf{x}, s_k(\mathbf{x}))$, determines the order in which the points are to be labeled. All points are initially labeled as uncommitted (or zero), and a committed (or non-zero) label can only be changed to another non-zero value. The label assignment procedure terminates when the objective function (23) can no longer be decreased.

6. Results and discussion

The algorithm has been tested on monocular indoor sequences. To reduce the computation afford, we assume

$\sigma_{b,k}^2(\mathbf{x}) = \sigma_{b,k}^2$ for every point at the step of Bayesian foreground detection in Section 4. Fig. 4 shows the segmentation results for the “aerobic” sequence. Fig. 4a shows four frames of the sequence. Using the same estimated background, the segmentation results of both simple background subtraction and our method are shown in Fig. 4b–d. Comparing with the results of simple background subtraction, the accuracy of object detection is greatly improved by the proposed approach. The moving cast shadows (the gray regions in Fig. 4c) are exactly removed from the foreground. The flickering background pixels that will be detected as foreground by simple background subtraction method are correctly classified by our algorithm. The camouflage at the neck makes the head almost separated from the body in Fig. 4b, while this effect is successfully overcome in Fig. 4d.

The comparison of the proposed method with two recent adaptive background subtraction techniques, background variation [1] and mixture of Gaussians [4], has also been investigated. The performance of the proposed technique, background variation (BV), and mixture of Gaussians (MG) is tested on a “laboratory” sequence. All the three methods are initialized using the first 50 frames of the sequence. A smoothing operation is applied on the detection results of BV and MG before comparison. The segmentation results are shown in Fig. 5. Fig. 5a shows four frames of the sequence. The manually segmented “ground truth” foreground images are shown in Fig. 5b. The segmentation results of BV, MG, and our technique are shown in Fig. 5c–e, respectively. Besides visual comparison, the results are also evaluated quantitatively in terms of the false negative number and rate (the number and portion of foreground pixels that are missed) and the false positive number and rate (the number and portion of non-foreground pixels that are marked as foreground) by comparing to the “ground truth” images. The errors for the four scenes in Fig. 5a are summarized in Table 1. It can be seen that moving shadows cast on the floor, wall, and table result in an increase of falsely detected foreground pixels in Fig. 5c and d. Large shadow attachments may cause failures in further analysis such as object recognition and tracking. Here cast shadows follow the movement of the person, so that they could not be learned by the background model as background changes. Moreover, there are a number of lights from the ceiling in the scene. Without an explicit shadow model, it is difficult to know which Gaussians in the mixture are produced by shadows.

Fig. 6 shows the segmentation results by the proposed method for another “laboratory” sequence. The open cabinet in the third and fourth images is classified as background after a period of background updating. However, it can be seen from Figs. 5e and 6b that erroneous segmentation sometimes takes place at boundary areas. The spatial constraint from the MRF formulation is relatively weak at object boundaries, so that errors are more likely to happen at these areas when the foreground has similar color as the

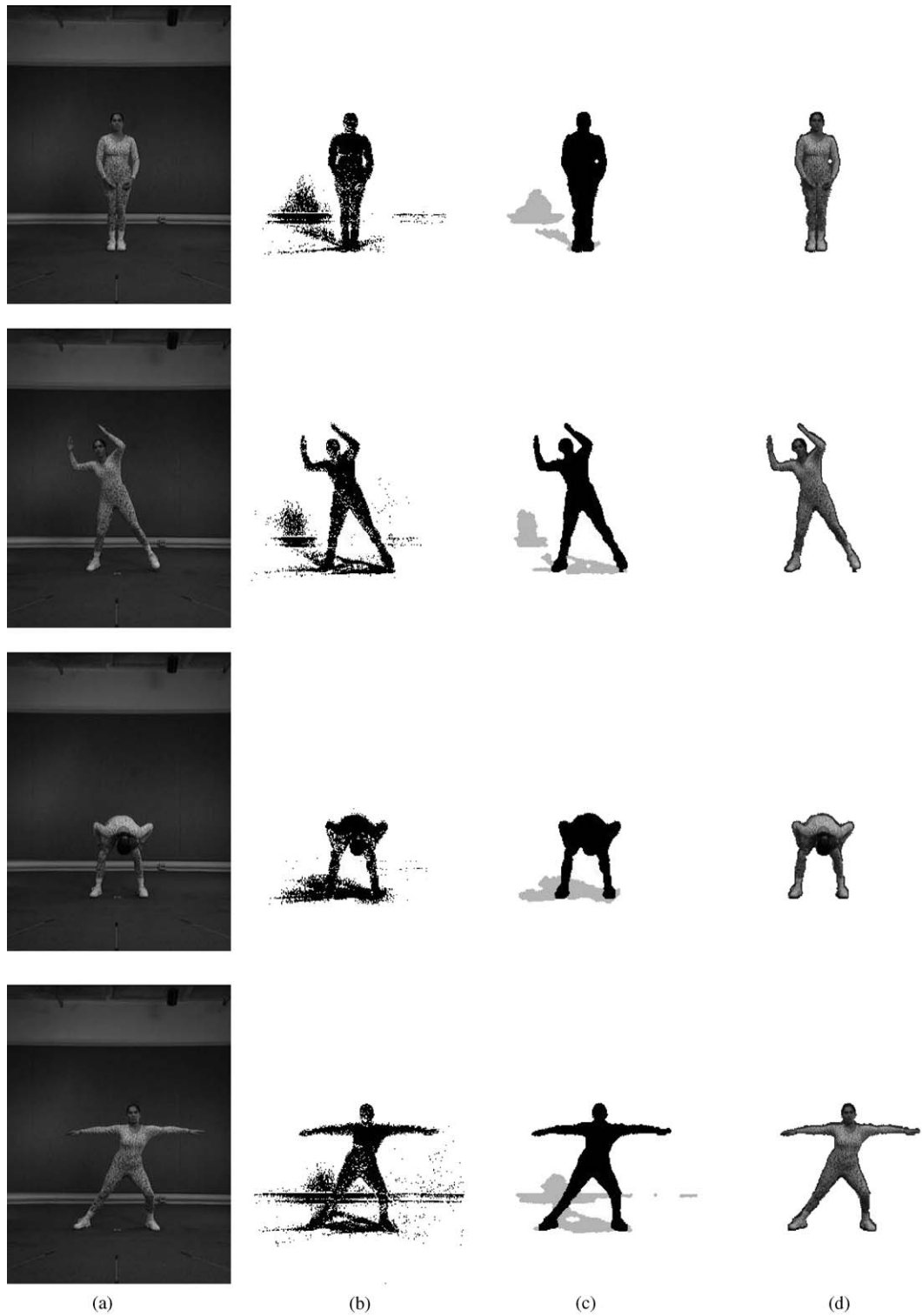


Fig. 4. (a) Frames of the “aerobic” sequence. (b) The segmentation results of simple background subtraction. (c) The segmentation results of the proposed algorithm. (d) The foreground detected by the proposed algorithm.

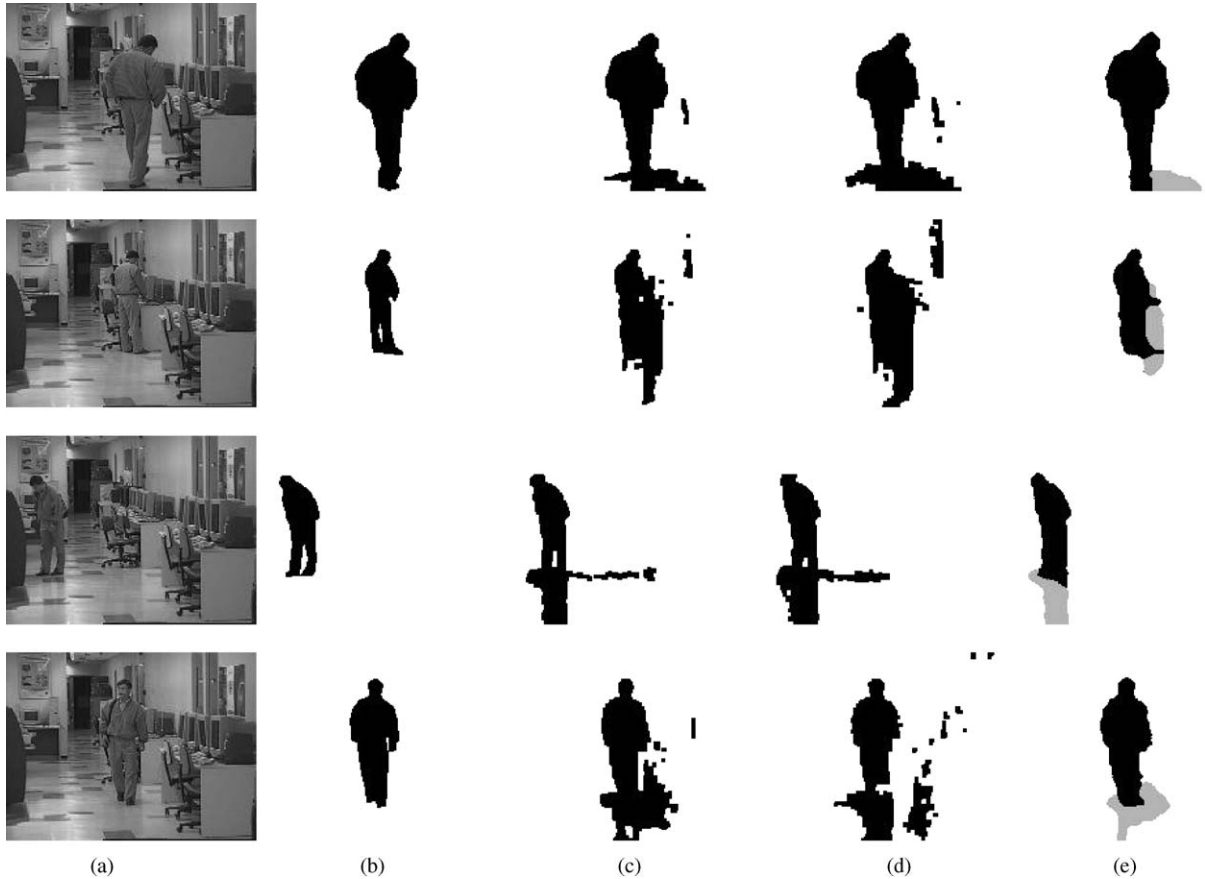


Fig. 5. (a) Frames of the “laboratory” sequence. (b) The “ground truth” foreground. (c) The segmentation results of BV. (d) The segmentation results of MG. (e) The segmentation results of the proposed algorithm.

Table 1
Quantitative evaluation of different methods

Method	False negative		False positive		Total errors
	Number	Rate (%)	Number	Rate (%)	
BV	254	3.0	5267	5.1	5521
MG	255	3.0	6904	6.7	7159
Proposed	178	2.1	961	0.9	1139

background or the pixel intensity under shadow is far from its mean.

During the segmentation process given in Section 4, the density of the image intensity at site \mathbf{x} is modeled as

$$\begin{aligned}
 & p(g_k(\mathbf{x})|\boldsymbol{\theta}_{b,k}(\mathbf{x})) \\
 &= \sum_{s_k(\mathbf{x})=1}^3 p(s_k(\mathbf{x}))p(g_k(\mathbf{x})|\boldsymbol{\theta}_{b,k}(\mathbf{x}), s_k(\mathbf{x})). \quad (30)
 \end{aligned}$$

Comparing this to the right side of (7a) in the case of $K = 3$, it can be found that uniform distribution is assumed for the foreground in (30), while Gaussian distribution is assumed in (7a). Since in the foreground there is no particular reason to prefer one value over any other, (30) could be thought as the improvement of (7a). However, the mixture of different kinds of distributions is much harder to estimate than the mixture of only Gaussians. Since foreground regions usually have large variances, from (9) we can see that such a difference will not make the backgrounding process in Section 3 to produce biased results.

7. Conclusion

In this paper we have presented an adaptive approach for foreground segmentation and shadow detection in monocular indoor image sequences. Graphical probabilistic models are employed in our approach. In our work, three sources of



Fig. 6. (a) Frames of another “laboratory” sequence. (b) The segmentation results of the proposed algorithm.

information are employed in object and shadow detection. The first is edge information, the difference images help locate changes in the scene. The second is spatial information, objects and shadows usually form continuous regions, and the third is temporal information, the models are updated from previous segmentation results.

Experimental results show that our method successfully deals with nonstationary background, camouflage and shadows in grayscale video sequences. Moreover, the algorithm can be easily implemented for color image sequences. How to further decrease the computation load of the optimization process and automatically determine all the parameters in our model is the topic of our future study.

Acknowledgements

The authors acknowledge Dr. Andrea Prati, Dr. James Davis, and Dr. Li-Yuan Li et al. for providing the test data on the website.

References

- [1] I. Haritaoglu, D. Harwood, L. Davis, W^4 : real-time surveillance of people and their activities, *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (2000) 809–830.
- [2] M. Seki, H. Fujiwara, K. Sumi, A robust subtraction method for changing background, *Proceedings of IEEE Workshop on Applications of Computer Vision*, 2000, pp. 207–213.
- [3] N. Friedman, S. Russell, Image segmentation in video sequences: a probabilistic approach, *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence 1997*, pp. 175–181.
- [4] C. Stauffer, W.E.L. Grimson, Learning patterns of activity using real-time tracking, *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (2000) 747–757.
- [5] A. Elgammal, R. Duraiswami, D. Harwood, L.S. Davis, Background and foreground modeling using nonparametric kernel density estimation for visual surveillance, *Proc. IEEE* 90 (2002) 1151–1163.
- [6] J. Rittscher, J. Kato, S. Joga, A. Blake, A probabilistic background model for tracking, *Proceedings of European Conference on Computer Vision*, vol. 2, 2000, pp. 336–350.
- [7] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, J.M. Buhmann, Topology free hidden Markov model: application to background modelling, *Proceedings of International Conference on Computer Vision*, 2001, pp. 294–301.
- [8] A. Prati, I. Mikic, M.M. Trivedi, R. Cucchiara, Detecting moving shadows: algorithms and evaluation, *IEEE Trans. Pattern Anal. Mach. Intell.* 25 (2003) 918–923.
- [9] G. Gordon, T. Darrell, M. Harville, J. Woodfill, Background estimation and removal based on range and color, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 1999, pp. 459–464.
- [10] S.J. McKenna, S. Jabri, Z. Duric, H. Wechsler, Tracking interacting people, *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 348–353.
- [11] C.R. Wren, A. Azarbayejani, T. Darrell, A.P. Pentland, Pfinder: real-time tracking of the human body, *IEEE Trans. Pattern Anal. Mach. Intell.* 19 (1997) 780–785.
- [12] S. Jabri, Z. Duric, H. Wechsler, A. Rosenfield, Detection and location of people in video images using adaptive fusion of color and edge information, *Proceedings of International Conference on Pattern Recognition*, vol. 4, 2000, pp. 627–630.
- [13] J. Stauder, R. Mech, J. Ostermann, Detection of moving cast shadows for object segmentation, *IEEE Trans. Multimedia*, vol. 1, 1999, pp. 65–76.
- [14] I. Mikic, P.C. Cosman, G.T. Kogut, M.M. Trivedi, Moving shadow and object detection in traffic scenes, *Proceedings of International Conference on Pattern Recognition*, vol. 1, 2000, pp. 321–324.
- [15] P.A. Flach, On the state of the art in machine learning: a personal review, *Artificial Intelligence* 131 (2001) 199–222.

- [16] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, S. Russell, Towards robust automatic traffic scene analysis in real-time, Proceedings of International Conference on Pattern Recognition, vol. 1, 1994, pp. 126–131.
- [17] N. Paragios, V. Ramesh, A MRF-based approach for real-time subway monitoring, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2001, pp. 1034–1040.
- [18] I. Patras, E.A. Hendriks, R.L. Lagendijk, Video segmentation by MAP labeling of watershed segments, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 326–332.
- [19] S.L. Dockstader, A.M. Tekalp, Multiple camera tracking of interacting and occluded human motion, Proc. IEEE 89 (2001) 1441–1455.
- [20] M. Harville, G. Gordon, J. Woodfill, Foreground segmentation using adaptive mixture models in color and depth, Proceedings of IEEE Workshop on Detection and Recognition of Events in Video, 2001, pp. 3–11.
- [21] F.V. Jensen, Bayesian Networks and Decision Graphs, Springer, Berlin, 2001.
- [22] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1984) 721–741.
- [23] A.M. Tekalp, Digital Video Processing, Prentice-Hall, Englewood Cliffs, 1995.
- [24] P.B. Chou, C.M. Brown, The theory and practice of Bayesian image labeling, Int. J. Comput. Vision 4 (1990) 185–210.
- [25] S.Z. Li, Markov Random Field Modeling in Image Analysis, Springer, Berlin, 2001.

About the Author—YANG WANG was born in China, 1976. He received his B.Eng. degree in Electronic Engineering and M.Sc. degree in Biomedical Engineering from Shanghai Jiao Tong University in 1998 and 2001, respectively. He obtained his Ph.D. degree in Computer Science from National University of Singapore in 2004. He was awarded the National Excellence Scholarship of China and the President's Graduate Fellowship of Singapore. Dr. Wang has published about 10 international journal and conference papers. His current research interests are in the area of Machine Intelligence and Computer Vision.

About the Author—TELE TAN is Senior Lecturer in the Division of Engineering, Science and Computing at the Curtin University of Technology, Western Australia, where he is affiliated to the Department of Computing, Department of Electrical and Computer Engineering and the Applied Physics Department. His research interests are in human motion analysis, security and surveillance, multi-modal system considerations and technology commercialization. Dr. Tan helped contribute to the original commercialization plan of a biometrics start-up company, XiD Technologies (<http://www.xidtech.com>) in late 2002. The biometrics software developed by the company was nominated for the 2004 World Technology Awards (software category) that was held in conjunction with the World Technology Summit 2004. He was made Technical Advisor to Miltrade Technologies in 2003 and was appointed International Reader with the Australian Research Council (ARC) in mid 2004.

About the Author—KIA-FOCK LOE is an Associate Professor in the Department of Computer Science at the National University of Singapore. He obtained his Ph.D. degree from the University of Tokyo. His current research interests are pattern recognition, computer vision, neural network, machine learning, and uncertainty reasoning.

About the Author—JIAN-KANG WU currently is principal Scientist, department manager of new initiatives, Institute for Infocomm Research (I2R), Singapore, which formally known as Kent Ridge Digital Labs (KRDL), and Institute of Systems Science (ISS), National University of Singapore. Dr. Wu received Bsc from the University of Science and Technology of China, and Ph.D. from Tokyo University. Prior to join ISS in 1992, he was a full professor in the University of Science and Technology of China, received 9 distinguished awards from the Ministry of Education and Ministry of Science of China and the Chinese Academy of Science. He also worked in universities in US, UK, Germany, France and Japan. Dr. Wu pioneered several researches in the area of visual information processing. This includes adaptive image coding in later 70s, object-oriented GIS in early 80s, face recognition system in 1992, content-based multimedia indexing and retrieval in early 90s, NeuroInformatics and PhysioInformatics recently. He initiated and led 3 large intentional collaboration projects in 90s. He is an author of 18 patents, 60+ journal publications and 5 books.