©2005 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

A Dynamic Hidden Markov Random Field Model for Foreground and Shadow Segmentation

Yang Wang[1,2], Kia-Fock Loe[1], Tele Tan[2], and Jian-Kang Wu[2]
[1]Dept. CS, National University of Singapore, Singapore 117543, yang.wang@computer.org, loekf@comp.nus.edu.sg
[2]Institute for Infocomm Research, Singapore 119613, {teletan, jiankang}@i2r.a-star.edu.sg

Abstract

This paper proposes a dynamic hidden Markov random field (DHMRF) model for foreground object and moving shadow segmentation in indoor video scenes. Given an image sequence, temporal dependencies of consecutive segmentation fields and spatial dependencies within each segmentation field are unified in the novel dynamic probabilistic model that combines the hidden Markov model (HMM) and the Markov random field (MRF). An efficient approximate filtering algorithm is derived for the DHMRF model to recursively estimate the segmentation field from the history of observed images. The foreground and shadow segmentation method integrates both intensity and edge information. Moreover, models of background, shadow, and edge information are updated adaptively for nonstationary background processes. Experimental results show that the proposed approach can accurately detect moving objects and their cast shadows even in monocular grayscale video sequences.

1. Introduction

Detecting moving objects in video sequences is very important in application areas such as visual surveillance, content-based video coding, and human computer interaction. When the video data is captured with a fixed camera, background subtraction is a commonly used technique to segment moving objects. The background model is constructed from observed images and foreground objects are identified if they differ significantly from the background. However, accurate foreground segmentation could be difficult due to the potential variability such as moving shadows cast by foreground objects, illumination or object changes in the background, and camouflage (i.e. similarity between appearances of foreground objects and the background) [2] [17] [25]. Besides local measurements such as depth and chromaticity [6] [8] [9] [14], constraints in temporal and spatial information from the video scene are very important to deal with the potential variability during the segmentation process.

Temporal or dynamic information is a fundamental element to handle the evolution of the scene. The background model can be adaptively updated from the recent history of observed images to handle nonstationary background processes (e.g. illumination changes). In addition, once a foreground point is detected, it will probably continue being in the foreground for some time. Linear prediction of background changes from recent observations can be performed by Kalman filter [12] or Wiener filter [24] to deal with dynamics in background processes. In the W^4 system [7], a bimodal background model is built for each site from order statistics of recent observed values. In [4], the pixel intensity is modeled by a mixture of three Gaussians (for moving object, shadow, and background respectively), and an incremental EM algorithm is used to learn the pixel model. In [22], the recent history of a pixel is modeled by a mixture of (usually three to five) Gaussians for nonstationary background processes. In [3], nonparametric kernel density estimation is employed for adaptive and robust background modeling. Moreover, a hidden Markov model (HMM) is used to impose the temporal continuity constraint on foreground and shadow detection for traffic surveillance [19]. A dynamical framework of topology free HMM capable of dealing with sudden or gradient illumination changes is also proposed in [23].

Spatial information is another essential element to understand the structure of the scene. Spatial variation information such as gradient (or edge) feature helps improve the reliability of structure change detection. In addition, contiguous points are likely to belong to the same background or foreground region. [10] classifies foreground versus background by adaptive fusion of color and edge information using confidence maps. [21] assumes that static edges in the background remain under shadow and that penumbras exist at the boundary of shadows. In [20], spatial cooccurrence of image variations at neighboring blocks is employed to improve the detection sensitivity of background subtraction. Moreover, spatial smooth constraint is imposed on moving object and shadow detection by propagating neighborhood information [15]. In [16], spatial interaction constraint is modeled by the Markov random field (MRF). In [11], a three dimensional MRF called spatio-temporal MRF involving two successive video frames is proposed for occlusion robust segmentation of traffic images.

A dynamic hidden Markov random field (DHMRF) model, which differs from the above mentioned models of

Proceedings of the Seventh IEEE Workshop on Applications of Computer Vision (WACV/MOTION'05)
0-7695-2271-8/05 $ 20.00 IEEE

IEEE COMPUTER SOCIETY

Authorized licensed use limited to: CURTIN UNIVERSITY OF TECHNOLOGY. Downloaded on May 20,2010 at 03:02:31 UTC from IEEE Xplore. Restrictions apply.

spatial or temporal constraints, is proposed in this paper for segmenting indoor foreground objects by background subtraction and shadow removal. Spatial and temporal dependencies in the segmentation process are unified in the dynamic probabilistic model (DHMRF) that combines the MRF and the HMM. A computationally efficient approximate filtering algorithm is derived for the DHMRF model to recursively estimate the segmentation field. Each pixel in the scene is classified as foreground, shadow, or background from the history of video images. The foreground segmentation method integrates both intensity and edge features, and it adaptively updates the models of background, shadow, and edge information. Experimental results show that the proposed approach robustly handles shadow and camouflage in nonstationary background scenes and improves the accuracy of foreground detection in monocular video sequences.

2. Dynamic hidden Markov random field

Given an image sequence $\{g_k\}$, the segmentation label for a point $\mathbf{x}$ within the kth image is denoted by $s_k(\mathbf{x})$. Label $s_k(\mathbf{x}) \in \{1, 2, \ldots, L\}$ assigns the point $\mathbf{x}$ to one of L (L equals 3 in this paper, see Section 3) classes at time k. Here $k \in \mathbf{N}$, $\mathbf{x} \in \mathbf{X}$, and $\mathbf{X}$ is the spatial domain of the video scene. The entire label field is expressed compactly as s_k. Spatial and temporal constraints in the segmentation process can be imposed through a dynamic model of statistical dependencies of neighboring sites.

2.1. DHMRF model

Given the observed data up to time k, the posterior probability distribution of the segmentation field s_k is modeled by a Markov random field [5] to formulate spatial dependencies. In the MRF model, if $N_{\mathbf{x}}$ is the neighborhood of a site $\mathbf{x}$, then the conditional distribution of a single label at $\mathbf{x}$ depends only on the labels within its neighborhood $N_{\mathbf{x}}$. According to the Hammersley-Clifford theorem, the probability is given by a Gibbs distribution that has the following form [13].

$$p(s_k \mid g_{1:k}) \propto \exp[-\sum_{c \in C} V_c(s_k(c) \mid g_{1:k})], \quad (1)$$

where $g_{1:k}$ denotes $\{g_1, g_2, \ldots, g_k\}$, C is the set of all cliques c, V_c is the clique potential function, and $s_k(c)$ denotes $\{s_k(\mathbf{x}) \mid \mathbf{x} \in c\}$. A clique is a set of pixels that are neighbors of each other, and the potential function V_c depends only on the points within clique c.

Only one-pixel and two-pixel cliques are used in our work. The one-pixel potential $V_{\mathbf{x}}(s_k(\mathbf{x}) \mid g_{1:k})$ reflects the information (or constraint) from the observation for a single site, and the two-pixel potential imposes the spatial constraint to form contiguous regions. To simplify the computation, the pairwise constraint is assumed to be independent of the observed images. Hence the two-point potential is written as $V_{\mathbf{x},\mathbf{y}}(s_k(\mathbf{x}), s_k(\mathbf{y}))$. The posterior distribution at time k becomes

$$p(s_k \mid g_{1:k}) \propto \exp\{-\sum_{\mathbf{x} \in \mathbf{X}} [V_{\mathbf{x}}(s_k(\mathbf{x}) \mid g_{1:k}) + \frac{1}{2} \sum_{\mathbf{y} \in N_{\mathbf{x}}} V_{\mathbf{x},\mathbf{y}}(s_k(\mathbf{x}), s_k(\mathbf{y}))]\}. \quad (2)$$

Spatial connectivity constraint can be imposed by the following two-pixel potential.

$$V_{\mathbf{x},\mathbf{y}}(s_k(\mathbf{x}) = i, s_k(\mathbf{y}) = j) \propto \frac{1}{\| \mathbf{x} - \mathbf{y} \|^2} (1 - \delta(i - j)), \quad (3)$$

where $1 \le i, j \le L$, $\|\cdot\|$ denotes the Euclidian distance, and $\delta(\cdot)$ is the Kronecker delta function. Thus two neighboring pixels are more likely to belong to the same class than to different classes. The spatial constraint becomes strong with decreasing distance between the neighboring sites.

The dynamic or temporal dependencies of consecutive segmentation fields are formulated by a hidden Markov model [18]. In the HMM, image g_k is the kth observation, and segmentation field s_k is the hidden state at time k. Therefore the state transition model $p(s_{k+1} \mid s_k)$ and the observation (or likelihood) model $p(g_k \mid s_k)$ for the HMM should be built for the entire scene.

The label field state transition probability $p(s_{k+1} \mid s_k)$ is modeled by a Markov random field defined on one-pixel and two-pixel cliques as well.

$$p(s_{k+1} \mid s_k) \propto \exp[-\sum_{c \in C} V_c(s_{k+1}(c) \mid s_k)]$$
$$= \exp\{-\sum_{\mathbf{x} \in \mathbf{X}} [V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) \mid s_k(M_{\mathbf{x}})) + \frac{1}{2} \sum_{\mathbf{y} \in N_{\mathbf{x}}} V_{\mathbf{x},\mathbf{y}}(s_{k+1}(\mathbf{x}), s_{k+1}(\mathbf{y}))]\}, \quad (4)$$

where $M_{\mathbf{x}}$ designates the set of sites in the kth image that impact on site $\mathbf{x}$ in the (k+1)th image. The one-pixel potential $V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) \mid s_k(M_{\mathbf{x}}))$ models the label state transition for a single site, and the two-pixel potential $V_{\mathbf{x},\mathbf{y}}(s_{k+1}(\mathbf{x}), s_{k+1}(\mathbf{y}))$ imposes the pairwise spatial constraint. It should be noted that $M_{\mathbf{x}}$ is not equivalent to the neighborhood $N_{\mathbf{x}}$. $M_{\mathbf{x}}$ and $N_{\mathbf{x}}$ may have different sizes. $\mathbf{x} \notin N_{\mathbf{x}}$ while $\mathbf{x} \in M_{\mathbf{x}}$ (e.g. see Figure 1). To distinguish them, $N_{\mathbf{x}}$ is called the spatial neighborhood, and $M_{\mathbf{x}}$ the temporal neighborhood.

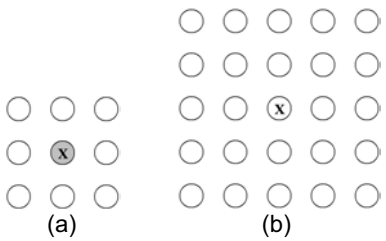


Figure 1. (a) The 8-pixel spatial neighborhood. (b) The 25-pixel temporal neighborhood.

Proceedings of the Seventh IEEE Workshop on Applications of Computer Vision (WACV/MOTION'05)
0-7695-2271-8/05 $ 20.00 IEEE

IEEE COMPUTER SOCIETY
Authorized licensed use limited to: CURTIN UNIVERSITY OF TECHNOLOGY. Downloaded on May 20,2010 at 03:02:31 UTC from IEEE Xplore. Restrictions apply.

Assuming conditional independence between spatially distinct observations, the observation model $p(g_k \mid s_k)$ is factorized as

$$p(g_k \mid s_k) = \prod_{\mathbf{x}\in\mathbf{X}} p(\mathbf{o}_k(\mathbf{x}) \mid s_k(\mathbf{x})), \quad (5)$$

where $\mathbf{o}_k(\mathbf{x})$ is the observation for site $\mathbf{x}$ that consists of locally measured information such as intensity and gradient features (see Section 3.2).

By (2), (4), and (5), spatial and temporal dependencies in the segmentation process are unified in a dynamic model that combines the MRF and the HMM. Therefore it is called the dynamic hidden Markov random field (DHMRF) model.

2.2. DHMRF filter

From a Bayesian perspective, the filtering algorithm is to recursively update the posterior distribution of the segmentation field. Given the potentials of the distribution $p(s_k \mid g_{1:k})$, the posterior $p(s_{k+1} \mid g_{1:k+1})$ at time k+1 can be efficiently approximated by a Markov random field with the following potential functions.

$$V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) = i \mid g_{1:k+1})$$
$$= -\ln\{\sum_j \exp[-\alpha_{i|j} - \lambda_k \sum_{\mathbf{y}\in M_{\mathbf{x}}} \frac{1}{|M_{\mathbf{y}}|} V_{\mathbf{y}}(s_k(\mathbf{y}) = j \mid g_{1:k})]\} -$$
$$\ln p(\mathbf{o}_{k+1}(\mathbf{x}) \mid s_{k+1}(\mathbf{x}) = i), \quad (6a)$$

$$V_{\mathbf{x},\mathbf{y}}(s_{k+1}(\mathbf{x}) = i, s_{k+1}(\mathbf{y}) = j) = \frac{\beta}{\|\mathbf{x}-\mathbf{y}\|^2}(1-\delta(i-j)), \quad (6b)$$

where $1 \le i, j \le L$, $|\cdot|$ denotes the size (number of points) of the set, $\alpha_{i|j}$ is the potential of state transition (from j to i) that imposes the temporal continuity constraint on segmentation label, λ_k and β weight the constraint from previous observations and the constraint of spatial connectivity respectively. The parameters are initialized and determined in Section 4.2. In the one-pixel potential (6a), the first term reflects the information from previously observed images for a single site $\mathbf{x}$, which is affected by its temporal neighborhood $M_{\mathbf{x}}$. The second term in (6a) reflects the information from the current observation. The two-pixel potential (6b) imposes the constraint from the spatial neighborhood.

3. Foreground and shadow segmentation

Given the video sequence, each pixel in the scene is to be classified as background, shadow, or foreground. For a site $\mathbf{x}$ in the kth frame, the segmentation label $s_k(\mathbf{x})$ equals 1 for a background pixel, 2 for shadow, and 3 for foreground. Here static shadows are considered to be part of the background.

3.1. Local observation

In order to segment the foreground, the system should first model the background and shadow information. Edge information also helps improve the reliability of detection.

Since indoor environments are relatively stable compared to outdoor scenes, we assume that each pixel in the background is of Gaussian distribution. At time k,

$$b_k(\mathbf{x}) = \mu_{b,k}(\mathbf{x}) + n_{b,k}(\mathbf{x}), \quad (7)$$

where random variable $b_k(\mathbf{x})$ is the intensity of a pixel $\mathbf{x}$ within the background, $\mu_{b,k}(\mathbf{x})$ is the intensity mean, and $n_{b,k}(\mathbf{x})$ is independent zero-mean Gaussian noise with variance $\sigma^2_{b,k}(\mathbf{x})$ at time k. Intensity means and variances in the background can be estimated from previous images (see Section 4.1).

Given the intensity of a background point, we use a linear model to describe the change of intensity for the same point when shadowed in the video frame. At time k,

$$g_k(\mathbf{x}) = a b_k(\mathbf{x}) + n_{s,k}(\mathbf{x}), \text{ if } s_k(\mathbf{x}) = 2, \quad (8)$$

where the coefficient $a \in [0,1]$, and $n_{s,k}(\mathbf{x})$ is independent zero-mean Gaussian noise with variance $\sigma^2_{s,k}(\mathbf{x})$ at time k. The shadow noise $n_{s,k}(\mathbf{x})$ models the deviation from the simple linear approximation in real visual environments, especially when the entire background scene is not flat. Since it is difficult to compute $\sigma^2_{s,k}(\mathbf{x})$ individually for every site $\mathbf{x}$ in the scene, we assume that $\sigma^2_{s,k}(\mathbf{x})$ equals $\rho^2\sigma^2_{b,k}(\mathbf{x})$, and that the shadow noise is independent of the background noise. Thus the intensity of a shadowed point is of Gaussian distribution with the following mean and variance.

$$E[g_k(\mathbf{x})] = a\mu_{b,k}(\mathbf{x}),$$
$$\mathrm{Var}[g_k(\mathbf{x})] = (a^2+\rho^2)\sigma^2_{b,k}(\mathbf{x}), \text{ if } s_k(\mathbf{x}) = 2. \quad (9)$$

Parameters a and ρ are manually determined. Their values depend on the visual environment, usually $0.5 \le a < 1$ and $0.5 \le \rho \le 1.5$ in indoor scenes.

The edge model is built by applying an edge operator to the scene. For a site $\mathbf{x}$, denote $\mathbf{x}_l$ and $\mathbf{x}_r$ as its two horizontally neighboring (left and right) points, $\mathbf{x}_u$ and $\mathbf{x}_d$ its two vertically neighboring (up and down) points. At time k, the image edge vector $\mathbf{e}_{g,k}(\mathbf{x})$ is denoted by $(e^h_{g,k}(\mathbf{x}), e^v_{g,k}(\mathbf{x}))^T$, where $e^h_{g,k}(\mathbf{x}) = g_k(\mathbf{x}_l) - g_k(\mathbf{x}_r)$ is the horizontal difference, and $e^v_{g,k}(\mathbf{x}) = g_k(\mathbf{x}_u) - g_k(\mathbf{x}_d)$ is the vertical difference. The entire image edge field is expressed as $\mathbf{e}_{g,k}$.

Similarly, we can model the edge information for the background. At time k, the background edge vector $\mathbf{e}_{b,k}(\mathbf{x})$ for a site $\mathbf{x}$ is denoted by $(e^h_{b,k}(\mathbf{x}), e^v_{b,k}(\mathbf{x}))^T$, where $e^h_{b,k}(\mathbf{x}) = b_k(\mathbf{x}_l) - b_k(\mathbf{x}_r)$ and $e^v_{b,k}(\mathbf{x}) = b_k(\mathbf{x}_u) - b_k(\mathbf{x}_d)$

IEEE COMPUTER SOCIETY
Authorized licensed use limited to: CURTIN UNIVERSITY OF TECHNOLOGY. Downloaded on May 20,2010 at 03:02:31 UTC from IEEE Xplore. Restrictions apply.

are the horizontal difference and the vertical difference respectively. It can be known from the background model that $\mathbf{e}_{b,k}(\mathbf{x})$ is of bivariate normal distribution. According to the independent background noise assumption, the corresponding mean $\boldsymbol{\mu}_{\mathbf{e},k}(\mathbf{x})$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{e},k}(\mathbf{x})$ of the distribution can be calculated from the intensity means and variances of the four neighboring points.

$$\boldsymbol{\mu}_{\mathbf{e},k}(\mathbf{x}) = (\mu_{b,k}(\mathbf{x}_l) - \mu_{b,k}(\mathbf{x}_r), \mu_{b,k}(\mathbf{x}_u) - \mu_{b,k}(\mathbf{x}_d))^T ,$$

$$\boldsymbol{\Sigma}_{\mathbf{e},k}(\mathbf{x}) = \begin{pmatrix} \sigma^2_{b,k}(\mathbf{x}_l) + \sigma^2_{b,k}(\mathbf{x}_r), 0 \\ 0, \sigma^2_{b,k}(\mathbf{x}_u) + \sigma^2_{b,k}(\mathbf{x}_d) \end{pmatrix}. \quad (10)$$

The edge model can be used to detect structure changes in the scene as edge features appear, vanish, or change orientation. Although other edge operators such as the Sobel operator can be applied as well, we use the above operator with a diagonal covariance matrix to simplify the computation.

3.2. Likelihood model

Since the image edge field $\mathbf{e}_{g,k}$ is totally determined by the image g_k, the observation (or likelihood) model $p(g_k \mid s_k)$ can be written as $p(g_k, \mathbf{e}_{g,k} \mid s_k)$. Then the factorization of the likelihood in (5) becomes

$$p(g_k \mid s_k) = p(g_k, \mathbf{e}_{g,k} \mid s_k)$$

$$= \prod_{\mathbf{x}\in\mathbf{X}} p(g_k(\mathbf{x}), \mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(\mathbf{x})), \quad (11)$$

where $\mathbf{o}_k(\mathbf{x})$ in (5) is replaced by $(g_k(\mathbf{x}), \mathbf{e}_{g,k}(\mathbf{x}))$ to integrate both intensity and edge features. Given the segmentation label, we assume that the image intensity and image edge are conditionally independent on each other at each site. Hence the local likelihood can be factorized as the product of intensity likelihood and edge likelihood.

$$p(g_k(\mathbf{x}), \mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(\mathbf{x}))$$

$$= p(g_k(\mathbf{x}) \mid s_k(\mathbf{x})) p(\mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(\mathbf{x})) . \quad (12)$$

When site $\mathbf{x}$ is in the background, the intensity likelihood can be calculated using the background model.

$$p(g_k(\mathbf{x}) \mid s_k(\mathbf{x}) = 1) = N(g_k(\mathbf{x}); \mu_{b,k}(\mathbf{x}), \sigma^2_{b,k}(\mathbf{x})) , \quad (13)$$

where $N(\mathbf{z}; \mathbf{m}, \boldsymbol{\Sigma})$ is a Gaussian distribution with argument $\mathbf{z}$, mean $\mathbf{m}$, and covariance $\boldsymbol{\Sigma}$.

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

$$p(g_k(\mathbf{x}) \mid s_k(\mathbf{x}) = 2)$$

$$= N(g_k(\mathbf{x}); a\mu_{b,k}(\mathbf{x}), (a^2 + \rho^2)\sigma^2_{b,k}(\mathbf{x})) . \quad (14)$$

When site $\mathbf{x}$ is in the foreground, the background has no influence on the pixel intensity information. Uniform distribution is assumed for the foreground pixel intensity. The conditional probability density becomes

$$p(g_k(\mathbf{x}) \mid s_k(\mathbf{x}) = 3) = \frac{1}{y_{\max}} . \quad (15)$$

Here $[0, y_{\max}]$ is the intensity range for a point in the scene.

For each point $\mathbf{x}$, denote the set of its four nearest neighboring points by $N'_{\mathbf{x}} = \{\mathbf{x}_l, \mathbf{x}_r, \mathbf{x}_u, \mathbf{x}_d\}$. Considering the spatial connectivity of the scene, we assume that the four neighboring points have the same segmentation label as $\mathbf{x}$. Thus the edge likelihood is approximated by

$$p(\mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(\mathbf{x}) = j) \approx p(\mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(N'_{\mathbf{x}}) = j) . \quad (16)$$

Similarly, when the area $N'_{\mathbf{x}}$ is in the background, the probability density can be computed by the edge model.

$$p(\mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(N'_{\mathbf{x}}) = 1)$$

$$= N(\mathbf{e}_{g,k}(\mathbf{x}); \boldsymbol{\mu}_{\mathbf{e},k}(\mathbf{x}), \boldsymbol{\Sigma}_{\mathbf{e},k}(\mathbf{x})) . \quad (17)$$

When the area $N'_{\mathbf{x}}$ is shadowed, the edge likelihood can be computed using the models in Section 3.1.

$$p(\mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(N'_{\mathbf{x}}) = 2)$$

$$= N(\mathbf{e}_{g,k}(\mathbf{x}); a\boldsymbol{\mu}_{\mathbf{e},k}(\mathbf{x}), (a^2 + \rho^2)\boldsymbol{\Sigma}_{\mathbf{e},k}(\mathbf{x})) . \quad (18)$$

When the area $N'_{\mathbf{x}}$ belongs to the foreground, we assume that the point intensity within the foreground is independent and identically distributed (i. i. d.). From (15), it can be known that

$$p(\mathbf{e}_{g,k}(\mathbf{x}) \mid s_k(N'_{\mathbf{x}}) = 3)$$

$$= p(e^h_{g,k}(\mathbf{x}) \mid s_k(N'_{\mathbf{x}}) = 3)\ p(e^v_{g,k}(\mathbf{x}) \mid s_k(N'_{\mathbf{x}}) = 3)$$

$$= \left(\frac{1}{y_{\max}} - \frac{|e^h_{g,k}(\mathbf{x})|}{y^2_{\max}}\right)\left(\frac{1}{y_{\max}} - \frac{|e^v_{g,k}(\mathbf{x})|}{y^2_{\max}}\right) . \quad (19)$$

3.3. Segmentation algorithm

Substitute $(g_{k+1}(\mathbf{x}), \mathbf{e}_{g,k+1}(\mathbf{x}))$ for $\mathbf{o}_{k+1}(\mathbf{x})$ in (6a) and combine the likelihood model in Section 3.2, then the one-pixel potential function for the segmentation field at time k+1 can be updated by the DHMRF filter.

$$V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) = i \mid g_{1:k+1})$$

$$= -\ln\{\sum_j \exp[-\alpha_{i|j} - \lambda_k \sum_{\mathbf{y}\in M_{\mathbf{x}}} \frac{1}{|M_{\mathbf{y}}|} V_{\mathbf{y}}(s_k(\mathbf{y}) = j \mid g_{1:k})]\} -$$

$$\ln(g_{k+1}(\mathbf{x}) \mid s_{k+1}(\mathbf{x}) = i) - \ln(\mathbf{e}_{g,k+1}(\mathbf{x}) \mid s_{k+1}(\mathbf{x}) = i) , \quad (20)$$

where $1 \le i, j \le 3$. Meanwhile the two-pixel potential $V_{\mathbf{x},\mathbf{y}}(s_{k+1}(\mathbf{x}), s_{k+1}(\mathbf{y}))$ can be calculated using (6b).

At time k+1, the MAP (maximum a posteriori) estimate of the segmentation field is computed as

$$\hat{s}_{k+1} = \arg\max_{s_{k+1}} p(s_{k+1} \mid g_{1:k+1})$$

$$= \arg\min_{s_{k+1}} \{\sum_{\mathbf{x}\in\mathbf{X}} [V_{\mathbf{x}}(s_{k+1}(\mathbf{x}) \mid g_{1:k+1}) +$$

$$\frac{1}{2} \sum_{\mathbf{y}\in N_{\mathbf{x}}} V_{\mathbf{x},\mathbf{y}}(s_{k+1}(\mathbf{x}), s_{k+1}(\mathbf{y}))]\} . \quad (21)$$

IEEE COMPUTER SOCIETY
Authorized licensed use limited to: CURTIN UNIVERSITY OF TECHNOLOGY. Downloaded on May 20,2010 at 03:02:31 UTC from IEEE Xplore. Restrictions apply.

4. Implementation

4.1. Background updating

For stationary background scenes, the intensity mean and variance of each background point can be estimated from a sequence of background images recorded at the beginning.

For nonstationary background scenes, the background updating process is based on the idea of Stauffer and Grimson [22]. The recent history of each pixel is modeled by a mixture of Gaussians. As parameters of the mixture model change, the Gaussian distribution that has the highest ratio of weight over variance is chosen as the background model. After the segmentation of an image, each pixel is checked to match the existing Gaussian distributions. For a matched Gaussian, its weight increases and the corresponding mean and variance are updated utilizing the pixel value. For unmatched distributions, the means and variances remain the same, while the weights should be renormalized. If none of the distributions match the pixel value, the distribution of the lowest weight is replaced with a Gaussian with the pixel value as its mean, initially low weight and high variance.

The main difference between the Gaussian mixture method and our approach in background updating is the definition of match. In [22], a Gaussian is matched if the pixel value is within 2.5 standard deviations of the distribution. In our work, if the point is classified as background by the segmentation algorithm (DHMRF filtering), then the Gaussian corresponding to the background model is matched, otherwise a Gaussian is matched if the value is within 2.5 standard deviations of the distribution. Thus the estimation by the DHMRF filter is employed in the updating process. Each time after background updating, the models of shadow and edge information can be updated by (9) and (10).

4.2. Parameters and optimization

In the one-pixel potential function (20), the potential of state transition is expressed as $\alpha_{i|j} \propto (1-\delta(i-j))$, so that segmentation labels for the same site are likely to remain the same at consecutive time instants. To balance the influence of the terms in (20), we assume that

$$\frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}\in\mathbf{X}}[\sum_j \lambda_k \sum_{\mathbf{y}\in M_\mathbf{x}} \frac{1}{|M_\mathbf{y}|} V_\mathbf{y}(s_k(\mathbf{y})=j \mid g_{1:k})] = \sum_j \alpha_{i|j} = \gamma. \quad (22)$$

Hence λ_k and $\alpha_{i|j}$ are estimated as

$$\lambda_k = \frac{\gamma}{\frac{1}{|\mathbf{X}|}\sum_{\mathbf{x}\in\mathbf{X}}\sum_j V_\mathbf{x}(s_k(\mathbf{x})=j \mid g_{1:k})},$$

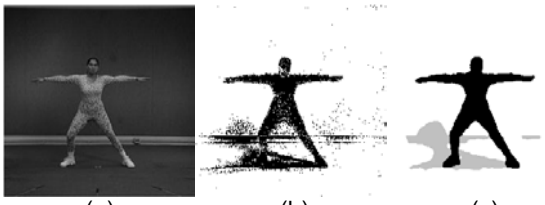
$$\alpha_{i|j} = \frac{1}{2}\gamma(1-\delta(i-j)),\ 1 \le i,j \le 3. \quad (23)$$

The parameters γ and β in (6b) are manually determined to reflect the importance of observed information and spatial connectivity respectively. Initially, the one-pixel potential $V_\mathbf{x}(s_0(\mathbf{x})=j) = \frac{\gamma}{3}$ for all $\mathbf{x}$ and j, and $\lambda_0 = 1$.

At each time, the MAP estimate is obtained by minimizing the objective function in (21). The objective function is nonconvex and does not have a unique minimum. Obviously, there is no simple method of performing the optimization. To arrive at a sub-optimal estimate, we use a local technique known as iterated conditional modes (ICM) [1]. The ICM algorithm employs the greedy strategy in iterative minimization. Initially, segmentation labels are set by maximizing the likelihood. Given the observed data and estimated labels of the latest iterative step, segmentation labels are sequentially updated by locally minimizing the objective function at each site.

5. Results and discussion

The proposed approach has been tested on monocular grayscale video sequences captured in different indoor environments. (For color images, they are first converted into grayscale ones.) Our C program can process about two 320×240 frames per second on a Pentium 4 2.8G Hz PC. Figure 2-3 show the segmentation results of two sequences with stationary background scenes, and Figure 4-5 show the segmentation results of two sequences with nonstationary background scenes. In Figure 4-5, our technique is compared to the Gaussian mixture (GM) method [22] and the method used in the W^4 system [7]. To save space, the figures show only part of the complete scenes. Unless otherwise stated, the segmentation results by our method are obtained using the 24-pixel spatial neighborhood and the 81-pixel temporal neighborhood.



(a) (b) (c)

Figure 2. (a) One frame of a sequence. (b) Segmentation result by simple background subtraction. (c) Segmentation result by the proposed method.

Figure 2 shows the segmentation results for one frame of the "aerobic" sequence using simple background subtraction and the proposed method. The gray regions in Figure 2c represent moving cast shadows. Compared to simple background subtraction, the proposed approach greatly improves the accuracy of foreground detection. The moving cast shadows attached to the woman in

IEEE COMPUTER SOCIETY
Authorized licensed use limited to: CURTIN UNIVERSITY OF TECHNOLOGY. Downloaded on May 20,2010 at 03:02:31 UTC from IEEE Xplore. Restrictions apply.

Figure 2b are exactly removed from the foreground in Figure 2c. The flickering pixels in the background and camouflage regions at the woman's neck and legs are erroneously detected in Figure 2b, while these problems are overcome by the proposed method.

(a.1) (a.2) (b.1) (b.2)

Figure 3. (a) Two frames of a sequence. (b) Segmentation results by the proposed method.

Figure 3 shows the segmentation results for two frames of the "room" sequence by the proposed method. Moving shadows cast at different locations of the wall and the floor are discriminated from the man in Figure 3b. When shadows are cast on multiple planes in the background scene, the noise term in the shadow model (8) ameliorates the linear approximation of intensity change under shadow.

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

Figure 4. (a) Two frames of a sequence. (b) Segmentation results by GM. (c) Segmentation results by W^4. (d) Segmentation results by the proposed method using the 4-pixel spatial neighborhood and the 9-pixel temporal neighborhood. (e) Segmentation results by the proposed method using the 24-pixel spatial neighborhood and the 81-pixel temporal neighborhood.

Figure 4 shows the segmentation results using GM, W^4, and the proposed method for two frames of the "laboratory" sequence with background object change. The open cabinet in the second image is classified as background in Figure 4b-4e by all the methods after a period of background updating. Figure 4d and 4e show the influence of neighborhood size. The camouflage regions and flickering areas in Figure 4d are corrected in Figure 4e by increasing spatio-temporal contextual constraints when the noise in the scene is heavy.

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

Figure 5. (a) Two frames of a sequence. (b) Segmentation results by GM. (c) Segmentation results by W^4. (d) Segmentation results by the proposed method without using edge information. (e) Segmentation results by the proposed method.

Figure 5 shows the segmentation results by GM, W^4, and the proposed method for two frames of another "laboratory" sequence with background illumination change. The illumination change in the second image caused by switching off part of the light is updated for the background in Figure 5b-5e by all the methods. Figure 5d and 5e show that the integration of edge information helps locate structure changes of the scene and improves the reliability of foreground detection.

Table 1. Quantitative evaluation of segmentation results.

	false negative	false positive
GM	2.9%	3.5%
W^4	5.0%	2.7%
proposed	3.2%	1.0%

The results are also evaluated quantitatively in terms of false negative rate (the portion of foreground pixels that are misclassified as non-foreground) and false positive rate (the portion of non-foreground pixels that are misclassified as foreground) by comparing to the manually segmented ground-truth foreground images. Before quantitative comparison, the segmentation results by the two other methods are smoothed to remove small erroneously detected areas. The average error rates for twenty frames of the two laboratory sequences (ten frames with different foreground object positions for each sequence) are summarized in Table 1. The moving shadows cast on the floor, wall, and table result in an increase of falsely detected foreground pixels (false positive) in Figure 4b-4c and 5b-5c. With an explicit shadow model, it is relatively easy for our approach to know which part of the pixel intensity distribution is likely to be produced by shadows. Moreover, both spatial and temporal constraints are employed in our approach. Hence the false positive rate is reduced by the proposed method with a tradeoff in relatively high computation load. On the other hand, in indoor scenes the intensity

Proceedings of the Seventh IEEE Workshop on Applications of Computer Vision (WACV/MOTION'05)
0-7695-2271-8/05 $ 20.00 IEEE

IEEE COMPUTER SOCIETY

Authorized licensed use limited to: CURTIN UNIVERSITY OF TECHNOLOGY. Downloaded on May 20,2010 at 03:02:31 UTC from IEEE Xplore. Restrictions apply.

variance of a point under shadow is usually greater than the variance of the same site in the background. Since the pixel intensity distribution of the foreground is assumed to be uniform, foreground regions darker than the background tend to be misclassified when the intensity variances under shadow are excessively large. This effect makes part of the man's arms erroneously detected as shadow in the first image of Figure 4e, and the false negative rate of our approach higher than that of the Gaussian mixture method.

6. Conclusion

There are two main contributions in this paper. First, we have proposed a dynamic hidden Markov random field (DHMRF) model that combines the HMM and the MRF for video sequences. Second, we have derived an efficient approximate DHMRF filtering algorithm and applied it to moving object and cast shadow detection in indoor scenes. The DHMRF model unifies the constraints of spatial connectivity and temporal continuity in the segmentation process. Objects and shadows usually form contiguous regions, and a point is likely to have the same segmentation label in consecutive frames. Two other kinds of spatial and temporal information are employed in our approach as well. The spatial gradient (or edge) information is integrated to help detect structure changes in the scene, and the recent history of observed images is used to adaptively update the models of background, shadow, and edge information.

The proposed approach does not require range or color data and performs robust foreground segmentation. Experimental results show that our method accurately distinguishes moving objects from their cast shadows in nonstationary background scenes. Our future study is to develop more accurate and efficient approximate filtering algorithms and automatically determine the parameters of the DHMRF model.

References

[1] J. Besag, "On the statistical analysis of dirty pictures," *J. R. Statist. Soc. B*, vol. 48, pp. 259-302, 1986.

[2] T. E. Boult, R. J. Micheals, X. Gao, and M. Eckmann, "Into the woods: Visual surveillance of noncooperative and camouflaged targets in complex outdoor settings," *Proc. IEEE*, vol. 89, pp. 1382-1402, 2001.

[3] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proc. IEEE*, vol. 90, pp. 1151-1163, 2002.

[4] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," *Proc. Conf. Uncertainty in Artificial Intelligence*, pp. 175-181, 1997.

[5] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 6, pp. 721-741, 1984.

[6] G. Gordon, T. Darrell, M. Harville, and J. Woodfill, "Background estimation and removal based on range and color," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 459-464, 1999.

[7] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: Real-time surveillance of people and their activities," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 22, pp. 809-830, 2000.

[8] T. Horprasert, D. Harwood, and L. S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," *Proc. FRAME-RATE Workshop*, 1999.

[9] Y. Ivanov, A. Bobick, and J. Liu, "Fast light independent background subtraction," *Int'l. J. Computer Vision*, vol. 37, pp. 199-207, 2000.

[10] S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld, "Detection and location of people in video images using adaptive fusion of color and edge information," *Proc. Int'l Conf. Pattern Recognition*, vol. 4, pp. 627-630, 2000.

[11] S. Kamijo, K. Ikeuchi, and M. Sakauchi, "Segmentations of spatio-temporal images by spatio-temporal Markov random field model," *Proc. EMMCVPR Workshop*, pp. 298-313, 2001.

[12] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russell, "Towards robust automatic traffic scene analysis in real-time," *Proc. Int'l Conf. Pattern Recognition*, vol. 1, pp. 126-131, 1994.

[13] S. Z. Li, *Markov Random Field Modeling in Image Analysis*, Springer-Verlag, 2001.

[14] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *Computer Vision and Image Understanding*, vol. 80, pp. 42-56, 2000.

[15] I. Mikic, P. C. Cosman, G. T. Kogut, and M. M. Trivedi, "Moving shadow and object detection in traffic scenes," *Proc. Int'l Conf. Pattern Recognition*, vol. 1, pp. 321-324, 2000.

[16] N. Paragios and V. Ramesh, "A MRF-based approach for real-time subway monitoring," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 1034-1040, 2001.

[17] A. Prati, I. Mikic, M. M. Trivedi, and R. Cucchiara, "Detecting moving shadows: Algorithms and evaluation," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 25, pp. 918-923, 2003.

[18] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257-286, 1989.

[19] J. Rittscher, J. Kato, S. Joga, and A. Blake, "A probabilistic background model for tracking," *Proc. European Conf. Computer Vision*, vol. 2, pp. 336-350, 2000.

[20] M. Seki, T. Wada, H. Fujiwara, and K. Sumi, "Background subtraction based on cooccurrence of image variations," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 65-72, 2003.

[21] J. Stauder, R. Mech, and J. Ostermann, "Detection of moving cast shadows for object segmentation," *IEEE Trans. Multimedia*, vol. 1, pp. 65-76, 1999.

[22] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 22, pp. 747-757, 2000.

[23] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, and J. M. Buhmann, "Topology free hidden Markov models: Application to background modeling," *Proc. Int'l Conf. Computer Vision*, vol. 1, pp. 294-301, 2001.

[24] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," *Proc. Int'l Conf. Computer Vision*, vol. 1, pp. 255-261, 1999.

[25] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 19, pp. 780-785, 1997.

IEEE COMPUTER SOCIETY

Authorized licensed use limited to: CURTIN UNIVERSITY OF TECHNOLOGY. Downloaded on May 20,2010 at 03:02:31 UTC from IEEE Xplore. Restrictions apply.