Spatiotemporal Saliency in Dynamic Scenes

Vijay Mahadevan, *Student Member*, *IEEE*, and
Nuno Vasconcelos, *Senior Member*, *IEEE*

Abstract—A spatiotemporal saliency algorithm based on a center-surround framework is proposed. The algorithm is inspired by biological mechanisms of motion-based perceptual grouping and extends a discriminant formulation of center-surround saliency previously proposed for static imagery. Under this formulation, the saliency of a location is equated to the power of a predefined set of features to discriminate between the visual stimuli in a center and a surround window, centered at that location. The features are spatiotemporal video patches and are modeled as dynamic textures, to achieve a principled joint characterization of the spatial and temporal components of saliency. The combination of discriminant center-surround saliency with the modeling power of dynamic textures yields a robust, versatile, and fully unsupervised spatiotemporal saliency algorithm, applicable to scenes with highly dynamic backgrounds and moving cameras. The related problem of background subtraction is treated as the complement of saliency detection, by classifying nonsalient (with respect to appearance and motion dynamics) points in the visual field as background. The algorithm is tested for background subtraction on challenging sequences, and shown to substantially outperform various state-of-the-art techniques. Quantitatively, its average error rate is almost half that of the closest competitor.

Index Terms—Spatiotemporal saliency, background subtraction, dynamic backgrounds, motion saliency, dynamic texture, discriminant center-surround architecture, video modeling.

1 Introduction

NATURAL scenes are usually composed of several dynamic entities. Foreground objects often move amid complicated backgrounds that are themselves moving, e.g., swaying trees or other objects such as a crowd, a flock of birds, moving water, waves, snow, rain, and smoke-filled environments. Even for static scenes, egomotion of the imaging sensor can cause a highly variable background. In the most extreme situations, egomotion and scene motion can combine to produce very complex background motion patterns. We refer to scenes with any of these types of variability as *dynamic scenes*. Since such scenes are plentiful in the natural world, successful discrimination between the background motions they induce and moving foreground objects, i.e., identifying regions that are *spatiotemporally salient*, is a strong survival advantage. Not surprisingly, biological visual systems have evolved to be extremely efficient in this task [3].

In computer vision, spatiotemporal saliency and the related task of background subtraction are commonly used as a preprocessing step for object and event detection [10], activity and gesture recognition [29], tracking [30], surveillance [12], and video retrieval [27]. Nevertheless, there has been little progress toward methods robust enough to handle the complexities of most dynamic scenes. Shortcomings of even the most advanced techniques include the requirements of

1. static cameras [12], [21], [25];
2. explicit [14], or approximate [23], compensation of camera motion;
3. foreground objects that move in a consistent direction (an assumption that we denote as *temporal coherence*) [6], [16], [28] or have faster variations in appearance than the background [24]; or
4. explicit background models.

These requirements are frequently unrealistic and particularly questionable when there is egomotion, e.g., a camera that tracks a moving object in a manner such that the latter has very small optical flow, or the background is dynamic. In addition, background learning requires a training set of "background-only" images [21], [25], [32] or batch processing (e.g., median filtering [10]) of a large number of video frames, which must be repeated for each scene and is difficult for dynamic scenes (where the background changes continuously).

We address these limitations through a novel spatiotemporal saliency paradigm, inspired by biological vision, where background subtraction is inherent to the deployment of visual attention. In particular, background subtraction is equated to the detection of salient motion, for which we propose a solution based on the *discriminant center-surround saliency hypothesis* [13]. Under this hypothesis, saliency is the result of optimal discrimination between center and surround stimuli at each location of the visual field. A set of visual features is collected from center and surround windows and the locations where the discrimination between the features of the two types can be performed with the smallest expected probability of error are declared as most salient. Background subtraction then reduces to ignoring the locations declared as nonsalient.

The center-surround formulation has various advantages over the traditional background subtraction procedures. First, there is no need to train or maintain models of the background. In fact, the proposed algorithm is completely unsupervised and does not require initialization with "background-only" frames. On the contrary, it is a *bottom-up* approach that can be equally applied to known and unknown scenes. Second, while a dynamic background is rarely homogeneous (e.g., different trees have different motion), spatial homogeneity usually holds locally. Hence, center-surround processing can be performed with much simpler probabilistic models (e.g., unimodal distributions versus mixtures) than those required to model the whole background. This simplifies parameter estimation. Third, since discriminant saliency only depends on the *relative disparity* between center and surround activity, it is invariant to camera motion. Finally, discriminant saliency is applicable to various problems, by simple modification of the features and probabilistic models used for center-surround discrimination. We adopt dynamic texture models, due to their versatility in modeling complex moving patterns and ability to replicate natural scene dynamics [11], [31]. This enables the proposed algorithm to account for joint saliency in motion and appearance, and makes it robust enough to handle complex dynamic backgrounds. Experimental results on a diverse collection of sequences show that the proposed algorithm substantially outperforms the current state of the art in background subtraction.

2 Biological Motivation

There is plentiful evidence that, in biological vision, bottom-up saliency is achieved through "center-surround" mechanisms tuned to detect stimuli that are distinct from stimuli in their surroundings [15], [19]. Extensive psychophysics experiments have shown that these mechanisms can be driven by a variety of features, including intensity, color, orientation, or motion, and *local feature contrast* plays a predominant role in the perception of saliency [22]. Fig. 1 shows some displays used in classical experiments designed to determine the role of feature contrast on judgments of motion saliency [22]. In one experiment, subjects were shown a display of moving dots such as that depicted in Fig. 1a (the videos are

• *The authors are with the Department of Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0407. E-mail: vmahadev@ucsd.edu, nuno@ece.ucsd.edu.*

Manuscript received 20 Feb. 2008; revised 12 Nov. 2008; accepted 30 Apr. 2009; published online 8 May 2009.
Recommended for acceptance by P. Perez.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-02-0110.
Digital Object Identifier no. 10.1109/TPAMI.2009.112.

0162-8828/10/$26.00 © 2010 IEEE Published by the IEEE Computer Society

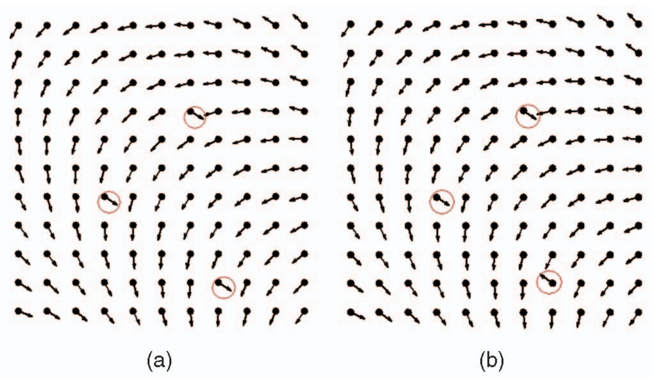


Fig. 1. Saliency perception due to local contrast [22]. Each panel shows a quiver plot of the stimuli (dots, whose direction of motion is indicated by arrows of length proportional to the speed of that motion). In (a), three targets which move in the same direction, among a field of distractors, are perceived as vertices of a moving triangle. This percept holds even if, as in (b), the direction of motion of one target is reversed. Video sequences of the two stimuli are available in [2].

available in [2]). While all dots (whose motion is indicated in the figure by arrows) were subject to motion different from that of their immediate neighbors, three (referred to as the *targets*, and indicated by circles in the figure) had *substantially larger motion contrast* than the others. The targets could be in different configurations, two of which are shown in the figure: 1) "similar" (Fig. 1a), where all three targets moved in the same direction and 2) "dissimilar" (Fig. 1b), where one target moved in a direction *different* than that of the other two. In all cases, subjects reported the percept of *pop-out* of a "moving triangle," with similar detection rates. While motion pop-out was already well established, these experiments showed that both motion saliency and the perceptual organization of the points into a triangle *do not depend on absolute quantities*, such as the direction of motion of the targets, how coherent their motion is, or the type of background motion. Instead, the coherent perception of the targets as a triangle, even when the vertex motions are incoherent and the background motion cannot be easily explained by a physical geometric transformation, suggests that both motion saliency and perceptual organization are driven by measurements of *local motion contrast*. Neurophysiological experiments on primates have also shown that neurons in the middle temporal visual area (MT) compute local motion contrast with center-surround mechanisms. It has, in fact, been hypothesized that such neurons underlie the perception of motion pop-out and figure-ground segmentation [4]. On the other hand, this evidence suggests that spatiotemporal saliency or background subtraction techniques which 1) rely on grouping of features by motion similarity to identify foreground objects or 2) require compensation of camera motion, will have difficulties to match the performance of biological systems.

From a computer vision point of view too, rooting saliency on measurements of local contrast appears to be a good idea. Note that, if motion contrast is defined as dissimilarity of optical flow, the saliency judgments are robust to egomotion. Furthermore, there is no need for a "global background model" or any type of training. Instead, saliency can be computed *efficiently* using purely local computations, and it *immediately adapts* to previously unseen environments. We will see in what follows that these properties still hold for *dynamic scenes*, under a more general definition of motion contrast.

3 Discriminant Center-Surround Spatiotemporal Saliency

The biological evidence of the previous section suggests the implementation of spatiotemporal saliency through local measurements of motion contrast. In this section, we propose an implementation based on the principle of *discriminant saliency* [13], with models of spatiotemoporal stimulus statistics that are suitable for dynamic scenes.

3.1 Mathematical Formulation

Discriminant saliency is defined with respect to two classes of stimuli: a class of *stimuli of interest* and a *background* or null hypothesis, consisting of stimuli that are not salient. The locations of the visual field that can be classified, with lowest expected probability of error, as containing stimuli of interest are denoted as salient. This is accomplished by setting up a binary classification problem between the stimuli of interest and the null hypothesis. The saliency of each location in the visual field is then equated to the discriminant power of the visual features extracted from that location in differentiating the two classes.

Formally, let $\mathcal{V}$ be a d-dimensional array representing the visual stimuli indexed by location vector $l \in L \subset \mathbb{R}^d$ and consider the responses of a predefined set of features $\boldsymbol{Y}$ (e.g., raw pixel values, Gabor or Fourier features), computed from $\mathcal{V}$ at all locations $l \in L$. A classification problem opposing two classes, of class label $C(l) \in \{0,1\}$, is posed at location l. Two windows are defined: a neighborhood $\mathcal{W}_l^1$ of l, which is denoted as *center*, and a surrounding annular window $\mathcal{W}_l^0$, which is denoted as the *surround*. The union of the two windows is denoted as the *total* window $\mathcal{W}_l = \mathcal{W}_l^0 \cup \mathcal{W}_l^1$. Let $\mathbf{y}^{(j)}$ be the vector of feature responses at location j. Features in the center $\{\mathbf{y}^{(j)} | j \in \mathcal{W}_l^1\}$ are drawn from the class of interest (or alternate hypothesis) $C(l) = 1$, with probability density $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|1)$. Features in the surround $\{\mathbf{y}^{(j)} | j \in \mathcal{W}_l^0\}$ are drawn from the null hypothesis $C(l) = 0$, with probability density $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|0)$. The saliency of location l, $S(l)$, is quantified by the mutual information between features $\boldsymbol{Y}$ and class label C,

$$S(l) = I_l(Y;C) = \sum_{c=0}^{1} \int_{\mathcal{Y}} p_{\boldsymbol{Y},C(l)}(\mathbf{y},c) \log \frac{p_{\boldsymbol{Y},C(l)}(\mathbf{y},c)}{p_{\boldsymbol{Y}}(\mathbf{y}) p_{C(l)}(c)} d\mathbf{y}, \tag{1}$$

where $y \in \mathcal{Y}$, the space of feature responses.

This is an approximation to the expected probability of success of center-surround classification (more precisely, one minus the Bayes error rate) [26]. A large $S(l)$ implies that center and surround have large disparity of feature responses, i.e., large *local feature contrast*, and can thus be discriminated with low probability of error. Conversely, locations where the classification has the smallest expected probability of error occur at maxima of $S(l)$. The function $S(l), l \in L$ is referred to as the *saliency map* of the stimuli $\mathcal{V}$ and can also be written as

$$S(l) = \sum_{c=0}^{1} p_{C(l)}(c) \mathrm{KL}\left(p_{\boldsymbol{Y}|C(l)}(\mathbf{y} \mid c) \| p_{\boldsymbol{Y}}(\mathbf{y})\right), \tag{2}$$

where $\mathrm{KL}\,(p\|q) = \int_{\mathcal{X}} p_{\boldsymbol{X}}(x) \log \frac{p_{\boldsymbol{X}}(x)}{q_{\boldsymbol{X}}(x)} dx$ is the Kullback-Leibler (KL) divergence between the probability distributions $p_{\boldsymbol{X}}(x)$ and $q_{\boldsymbol{X}}(x)$ [20].

3.2 Modeling Spatiotemporal Stimulus Statistics

Under this formulation, spatiotemporal saliency for highly dynamic scenes only requires the use, in (2), of probability models $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|c)$ that account for the variability of such scenes. We adopt the dynamic texture (DT) model of [11], due to its ability to account for this variability, while jointly modeling the spatiotemporal characteristics of the visual stimulus. A dynamic texture is an autoregressive model that represents the appearance of the stimulus $\mathbf{y}_t \in \mathbb{R}^m$ (the pixels of the two-dimensional visual stimulus are represented as a column vector of length m), observed at time t, as a linear function of a hidden state process $\mathbf{x}_t \in \mathbb{R}^n$ ($n \ll m$) subject to Gaussian observation noise. The state and appearance processes form a linear dynamical system (LDS)

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{v}_t, \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \mathbf{w}_t, \end{aligned} \tag{3}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is the state transition matrix, $\mathbf{C} \in \mathbb{R}^{m \times n}$ the observation matrix, and $\mathbf{v}_t \sim_{iid} \mathcal{N}(0, \mathbf{Q})$ and $\mathbf{w}_t \sim_{iid} \mathcal{N}(0, \mathbf{R})$ are Gaussian state and observation noise processes, respectively. The initial state is assumed to be distributed as $\mathbf{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{S}_1)$, and the model is parameterized by $\boldsymbol{\Theta} = (\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}, \boldsymbol{\mu}_1, \mathbf{S}_1)$.

3.3 Probability Distributions

Since the states of a DT form a Markov process with Gaussian conditional probability for $\mathbf{x}_t$ given $\mathbf{x}_{t-1}$ (for any t), and Gaussian initial conditions, the joint distribution of the state sequence, $\mathbf{x}(\tau) = [\mathbf{x}_1^T \cdots \mathbf{x}_\tau^T]^T$, is also Gaussian [8]

$$p_{\boldsymbol{X}}(\mathbf{x}(\tau)) \sim \mathcal{N}(\boldsymbol{\mu}(\tau), \boldsymbol{\Sigma}(\tau)), \tag{4}$$

with parameters defined by the recursions

$$\boldsymbol{\mu}(t) = \begin{bmatrix} \boldsymbol{\mu}(t-1) \\ \boldsymbol{\mu}_t \end{bmatrix}, \quad \boldsymbol{\Sigma}(t) = \begin{bmatrix} \boldsymbol{\Sigma}(t-1) & \boldsymbol{\Upsilon}^T(t) \\ \boldsymbol{\Upsilon}(t) & \mathbf{S}_t \end{bmatrix}, \tag{5}$$

where $\boldsymbol{\Upsilon}(t)$ is the cross-covariance between the state at time t and the sequence of past states, and

$$\boldsymbol{\mu}_t = \mathbf{A}\boldsymbol{\mu}_{t-1}, \quad \mathbf{S}_t = \mathbf{A}\mathbf{S}_{t-1}\mathbf{A}^T + \mathbf{Q}, \tag{6}$$

$$\boldsymbol{\Upsilon}(t) = [\mathbf{A}\boldsymbol{\Upsilon}(t-1) \quad \mathbf{A}\mathbf{S}_{t-1}], \tag{7}$$

for $t \in [2, \tau]$, with $\boldsymbol{\mu}(1) = \boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}(1) = \mathbf{S}_1$, and $\boldsymbol{\Upsilon}(2) = \mathbf{A}\mathbf{S}_1$. Similarly, the sequence of observations $\mathbf{y}(\tau) = [\mathbf{y}_1^T \cdots \mathbf{y}_\tau^T]^T$ has joint distribution

$$p_{\boldsymbol{Y}}(\mathbf{y}(\tau)) \sim \mathcal{N}(\gamma(\tau), \Phi(\tau)), \tag{8}$$

with parameters defined by the recursions,

$$\gamma(t) = \begin{bmatrix} \gamma(t-1) \\ \mathbf{C}\boldsymbol{\mu}_t \end{bmatrix}, \quad \boldsymbol{\Phi}(t) = \begin{bmatrix} \boldsymbol{\Phi}(t-1) & \zeta^T(t)\mathbf{C}^T \\ \mathbf{C}\zeta(t) & \mathbf{C}\mathbf{S}_\mathbf{t}\mathbf{C}^T + \mathbf{R} \end{bmatrix}, \tag{9}$$

$$\zeta(t) = [\mathbf{A}\zeta(t-1) \quad \mathbf{A}\mathbf{S}_{t-1}\mathbf{C}^T], \tag{10}$$

for $t \in [2, \tau]$, where $\mathbf{C}\zeta(t)$ is the cross-covariance between the observation at t and the past observation sequence. The initial conditions are $\gamma(1) = \mathbf{C}\boldsymbol{\mu}_1$, $\boldsymbol{\Phi}(1) = \mathbf{C}\mathbf{S}_1\mathbf{C}^T + \mathbf{R}$, and $\zeta(2) = \mathbf{A}\mathbf{S}_1\mathbf{C}^T$. Using the parameter estimates obtained from a collection of spatiotemporal patches extracted from the center and surround windows with the method of [11] in (4) and (8) produces the probability distributions required by (2).

3.4 KL Divergence between DTs

To evaluate the KL divergences of (2), let $p_{Y|C(l)}(\mathbf{y}(\tau)|c) \sim \mathcal{N}(\gamma_c(\tau), \Phi_c(\tau)), c \in \{0, 1\}$ be the class-conditional distributions of a sequence of τ frames under two DTs parameterized by $\boldsymbol{\Theta}_c(l)$, $c \in \{0,1\}$, respectively, and $p_{\boldsymbol{Y}}(\mathbf{y}(\tau)) \sim \mathcal{N}(\gamma(\tau), \boldsymbol{\Phi}(\tau))$ the marginal distribution parameterized by $\boldsymbol{\Theta}(l)$. Since all distributions are Gaussian, the KL divergence has closed-form [9]

$$\begin{aligned} &\mathrm{KL}\ (p_{\boldsymbol{Y}|C(l)}(\mathbf{y}(\tau)|c) \| p_{\boldsymbol{Y}}(\mathbf{y}(\tau))) \\ &= \frac{1}{2}\Bigg[\log\frac{|\boldsymbol{\Phi}(\tau)|}{|\boldsymbol{\Phi}_c(\tau)|} + \mathrm{tr}(\boldsymbol{\Phi}(\tau)^{-1}\boldsymbol{\Phi}_c(\tau)) \\ &\quad + \|\gamma_c(\tau) - \gamma(\tau)\|^2_{\boldsymbol{\Phi}(\tau)} - m\tau\Bigg], \end{aligned} \tag{11}$$

where m is the number of pixels in each frame, $\|\mathbf{z}\|_\mathbf{A} = \mathbf{z}^T\mathbf{A}^{-1}\mathbf{z}$ the Mahalanobis norm of $\mathbf{z}$ with respect to covariance $\mathbf{A}$, and $|\mathbf{A}|$ the determinant of $\mathbf{A}$. Direct evaluation of (11) is intractable since the matrices $\boldsymbol{\Phi}(\tau), \boldsymbol{\Phi}_c(\tau)$ have size $m\tau \times m\tau$. It is, however, possible to rewrite all terms in a recursive form that only depends on $n\tau \times n\tau$ matrices (recall that n is the dimension of the state space and $n \ll m$). The recursions are derived in full generality in [7]. Here, we only summarize the recursion associated with the case where the image noise is independently distributed, i.e., where the covariances $\mathbf{R}$, $\mathbf{R}_c$ of $\mathbf{w}_t$ in (3) are diagonal, i.e., $\mathbf{R} = \sigma^2\mathbf{I}$, $\mathbf{R}_c = \sigma_c^2\mathbf{I}$.

From the recursive definitions of $\gamma(\tau)$, $\boldsymbol{\mu}(\tau)$, $\boldsymbol{\Phi}(\tau)$, and $\boldsymbol{\Sigma}(\tau)$ in (9) and (5), it follows that the Mahalanobis term of (11) can be written as

$$\|\gamma_c(\tau) - \gamma(\tau)\|^2_{\boldsymbol{\Phi}(\tau)} = \|\gamma_c(\tau-1) - \gamma(\tau-1)\|^2_{\boldsymbol{\Phi}(\tau-1)} + \|\mathbf{z}_c(\tau)\|^2_{\hat{\boldsymbol{\Phi}}}, \tag{12}$$

where the update term is

$$\|\mathbf{z}_c(\tau)\|^2_{\hat{\boldsymbol{\Phi}}} = \frac{1}{\sigma^2}\|\mathbf{z}_c(\tau)\|^2 - \frac{1}{\sigma^4}\mathbf{z}_c^T(\tau)\mathbf{C}\boldsymbol{\Gamma}^{-1}(\tau)\mathbf{C}^T\mathbf{z}_c(\tau),$$

and

$$\mathbf{z}_c(\tau) = \frac{1}{\sigma^2}\mathbf{C}\boldsymbol{\Upsilon}(\tau)\left(\mathbf{I} - \frac{1}{\sigma^2}\boldsymbol{\beta}(\tau-1)\right)\boldsymbol{\nu}_c(\tau-1) - \gamma_{c,\tau} + \gamma_\tau, \tag{13}$$

$$\boldsymbol{\nu}_c(\tau) = \begin{bmatrix} \boldsymbol{\nu}_c(\tau-1) \\ \mathbf{C}^T\mathbf{C}_c\boldsymbol{\mu}_{c,\tau} - \boldsymbol{\mu}_\tau \end{bmatrix}, \ \boldsymbol{\nu}_c(1) = \mathbf{C}^T\mathbf{C}_c\boldsymbol{\mu}_{c,1} - \boldsymbol{\mu}_1, \tag{14}$$

$$\boldsymbol{\Gamma}(\tau) = \left[\mathbf{S}_\tau - \frac{1}{\sigma^2}\boldsymbol{\Upsilon}(\tau)\left(\mathbf{I} - \frac{1}{\sigma^2}\beta(\tau-1)\right)\boldsymbol{\Upsilon}^T(\tau)\right]^{-1} + \frac{1}{\sigma^2}\mathbf{I}, \tag{15}$$

$$\boldsymbol{\beta}(\tau) = \begin{bmatrix} \mathbf{H}^{-1}(\tau) & \mathbf{H}^{-1}(\tau)\mathbf{G}^T(\tau) \\ \mathbf{G}(\tau)\mathbf{H}^{-1}(\tau) & \boldsymbol{\beta}(\tau-1) + \mathbf{G}(\tau)\mathbf{H}^{-1}(\tau)\mathbf{G}^T(\tau) \end{bmatrix}, \tag{16}$$

$$\mathbf{G}(\tau) = -\begin{bmatrix} \mathbf{H}^{-1}(\tau-1)\boldsymbol{\Omega} \\ \mathbf{G}(\tau-1)\mathbf{H}^{-1}(\tau-1)\boldsymbol{\Omega} \end{bmatrix}, \tag{17}$$

with

$$\begin{aligned} \boldsymbol{\Omega} = -\mathbf{Q}^{-1}\mathbf{A}, \ \Xi = \mathbf{S}_1^{-1} + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}, \\ \mathbf{H}(\tau) = \Xi + \frac{1}{\sigma^2}\mathbf{I} - \boldsymbol{\Omega}^T\mathbf{H}^{-1}(\tau-1)\boldsymbol{\Omega}, \end{aligned}$$

and initial conditions

$$\begin{aligned} &\mathbf{G}(2) = -\boldsymbol{\beta}(1)\boldsymbol{\Omega}, \mathbf{H}(2) = \Xi + \frac{1}{\sigma^2}\mathbf{I} - \boldsymbol{\Omega}^T\boldsymbol{\beta}(1)\boldsymbol{\Omega}, \text{ and} \\ &\boldsymbol{\beta}(1) = \left[\mathbf{S}_1^{-1} + \frac{1}{\sigma^2}\mathbf{I}\right]^{-1}. \end{aligned}$$

This computation requires the inverse of

$$\boldsymbol{\Gamma}(\tau), \left[\mathbf{S}_\tau - \frac{1}{\sigma^2}\boldsymbol{\Upsilon}(\tau)\left(\mathbf{I} - \frac{1}{\sigma^2}\boldsymbol{\beta}(\tau-1)\right)\boldsymbol{\Upsilon}^T(\tau)\right],$$

$\mathbf{H}(\tau)$, and $[\mathbf{S}_1^{-1} + \frac{1}{\sigma^2}\mathbf{I}]$, which are all matrices of size $n \times n$.

The trace term has recursion,

$$\mathrm{tr}\left[\boldsymbol{\Phi}^{-1}(\tau)\boldsymbol{\Phi}_c(\tau)\right] = \omega_c(\tau) - \mathrm{tr}[\boldsymbol{\beta}(\tau)\boldsymbol{\Psi}_c(\tau)],$$

with

$$\begin{aligned} \omega_c(\tau) = &\frac{1}{\sigma^2}\mathrm{tr}[\mathbf{S}_{c,\tau}] + m\frac{\sigma_c^2}{\sigma^2} - \frac{\sigma_c^2}{\sigma^4}\mathrm{tr}[\mathbf{H}^{-1}(\tau)] \\ &- \frac{\sigma_c^2}{\sigma^4}\mathrm{tr}[\mathbf{H}^{-1}(\tau)\mathbf{G}^T(\tau)\mathbf{G}(\tau)] + \omega_c(\tau-1), \end{aligned} \tag{18}$$

$$\boldsymbol{\Psi}_c(\tau) = \begin{bmatrix} \boldsymbol{\Psi}_c(\tau-1) & \boldsymbol{\xi}_c^T(\tau)\mathbf{T}_c^T \\ \mathbf{T}_c\xi_c(\tau) & \frac{1}{\sigma^4}\mathbf{T}_c\mathbf{S}_{c,\tau}\mathbf{T}_c^T \end{bmatrix}, \tag{19}$$

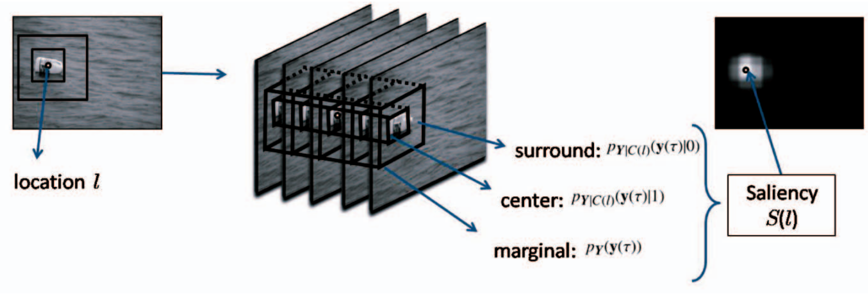

location l
surround: $p_{Y|C(l)}(\mathbf{y}(\tau)|0)$
center: $p_{Y|C(l)}(\mathbf{y}(\tau)|1)$
marginal: $p_{Y}(\mathbf{y}(\tau))$
Saliency
$S(l)$

Fig. 2. Illustration of the center and surround windows used to compute the saliency of location l. Conditional distributions are learned from the center and surround window, while the marginal distribution is learned from the total window. The saliency $S(l)$ is computed with (2).

$$\boldsymbol{\xi}_c(\tau) = \frac{1}{\sigma^4}\left[\mathbf{A}_c\boldsymbol{\xi}_c(\tau-1) \quad \mathbf{A}_c\mathbf{S}_{c,\tau-1}\mathbf{T}_c^T\right], \tag{20}$$

where $\mathbf{T}_c = \mathbf{C}^T\mathbf{C}_c$, and the initial conditions are $\omega_c(1) = \frac{1}{\sigma^2}\mathrm{tr}[\mathbf{S}_{c,1}] + m\frac{\sigma_c^2}{\sigma^2} - \frac{\sigma_c^2}{\sigma^4}\mathrm{tr}[\beta(1)]$, $\boldsymbol{\Psi}_c(1) = \frac{1}{\sigma^4}\mathbf{T}_c\mathbf{S}_{c,1}\mathbf{T}_c^T$, and $\boldsymbol{\xi}_c(2) = \frac{1}{\sigma^4}\mathbf{A}_c S_{c,1}\mathbf{T}_c^T$.

Finally, the determinant of $\boldsymbol{\Phi}(\tau)$ is given by

$$\log|\boldsymbol{\Phi}(\tau)| = \log|\boldsymbol{\Phi}(\tau-1)| + \sum_{k=1}^{n}\log\left(\frac{\lambda^{(k)}}{\sigma^2}+1\right) + m\log\sigma^2, \tag{21}$$

where $\lambda^{(k)}$ is the kth eigenvalue of

$$\left[\mathbf{S}_\tau - \frac{1}{\sigma^2}\boldsymbol{\Upsilon}(\tau)\left(\mathbf{I} - \frac{1}{\sigma^2}\beta(\tau-1)\right)\boldsymbol{\Upsilon}^T(\tau)\right],$$

an $n \times n$ matrix. The determinant of $\boldsymbol{\Phi}_c(\tau)$ is computed in a similar manner.

3.5 Background Subtraction

Background pixels are identified by computing the saliency $S(l)$ of each location l. Center and surround windows are defined at l, and a collection of spatiotemporal patches extracted from each window. Prior probabilities for both classes are assumed equal, and DT parameters are learned from center, surround, and total windows, to obtain the densities $p_{\mathbf{Y}|C(l)}(\mathbf{y}(\tau)|1)$, $p_{\mathbf{Y}|C(l)}(\mathbf{y}(\tau)|0)$, and $p_{\mathbf{Y}}(\mathbf{y}(\tau))$, respectively. $S(l)$ is finally computed with (2), using the recursions of (12)-(21). The procedure is summarized by Algorithm 1, and illustrated in Fig. 2. All locations with saliency below a threshold are assigned to the background.

Algorithm 1. Computing Discriminant Center Surround Motion Saliency

```
Input: Given video $\mathcal{V}$ indexed by location vector $l \in L \subset \mathbb{R}^3$,
  state-space dimension $n$, center window size $n_c$, patch size $n_p$,
  temporal window $\tau$.
for $l \in L$ do
    Identify center window $\mathcal{W}_l^1$ of size $n_c \times n_c \times \tau$ and
    surround window $\mathcal{W}_l^0$ of size $6n_c \times 6n_c \times \tau$ around $l$.
    Use all overlapping patches of size $n_p \times n_p \times \tau$ in $\mathcal{W}_l^1$, $\mathcal{W}_l^0$,
    and $\mathcal{W}_l^1 \cup \mathcal{W}_l^0$, to learn the dynamic texture parameters of
    the center $\boldsymbol{\Theta}_1(l)$, surround $\boldsymbol{\Theta}_0(l)$, and total $\boldsymbol{\Theta}(l)$
    distributions, using the method of [11].
    Compute the mutual information, $S(l)$, between
    class-conditional and total densities (2), using the recursive
    implementation of (11) given by (12)-(21).
end for
Output: Saliency map for $S(l), l \in L$
```

4 Experimental Evaluation

To evaluate background subtraction performance, Algorithm 1 was tested on video sequences collected from standard test sets (e.g., [5]) and the Web. Some of these sequences are representative of the classic application scenarios for background subtraction, e.g., a static camera that monitors a distant pedestrian walkway or a crowded highway. Others involve highly dynamic backgrounds (consisting of water, smoke, fire, or even a flock of birds), significant camera motion, or both. Representative frames from some of the sequences in the latter class are shown in Fig. 3. All sequences are available in [2].

4.1 Comparison to Previous Methods

To compare the performance of the proposed algorithm (denoted in short as DiscSal) with existing methods, we selected four representatives of the current state of the art in background subtraction—the modified Gaussian mixture model (GMM) of [1], [32], the nonparametric kernel density estimator (KDE) of [12], the linear dynamical model of Monnet et al. [21], and the "surprise" model proposed by Itti and Baldi [17], [18]. The original implementation of Monnet et al. [21] is not publicly available, and the algorithm requires explicit training with background frames. Since training data were not available for the sequences considered, we implemented an adaptive version, where the autoregressive model parameters were estimated from the 20 frames preceding the location under consideration. The higher adaptiveness of this version allows for a fairer comparison to saliency-based background subtraction.

The sequences were converted to grayscale, and saliency computed at all pixel locations. At each location, the center window occupied 16×16 pixels and spanned 11 frames—5 past, current, and 5 future ($n_c = 16$, $\tau = 11$). A causal version of Algorithm 1 (denoted DiscSal-Causal) was also implemented by considering only the current and 10 past frames. In all cases, the surround window was set to six times the size of the center (i.e., $96 \times 96 \times 11$). DTs with a 10-dimensional state space, patch dimension $n_p = 8$, and temporal dimension $\tau = 11$, were learned using overlapping $8 \times 8 \times 11$ patches from the center and surround windows. Saliency maps obtained with DiscSal, Surprise, KDE, Monnet, and GMM are shown in Fig. 3 (since the results for DiscSal and the causal version, DiscSal-Causal, were very similar, we omit the latter). Videos of the maps obtained for all sequences are available in [2]. The proposed algorithm clearly has the best performance, detecting the foreground motion and ignoring the complex moving background almost entirely. For all other methods, foreground detection is very noisy, and does not adapt well to the fast background dynamics. As a result, the saliency maps contain substantial energy in background regions, sometimes missing the foreground objects completely.

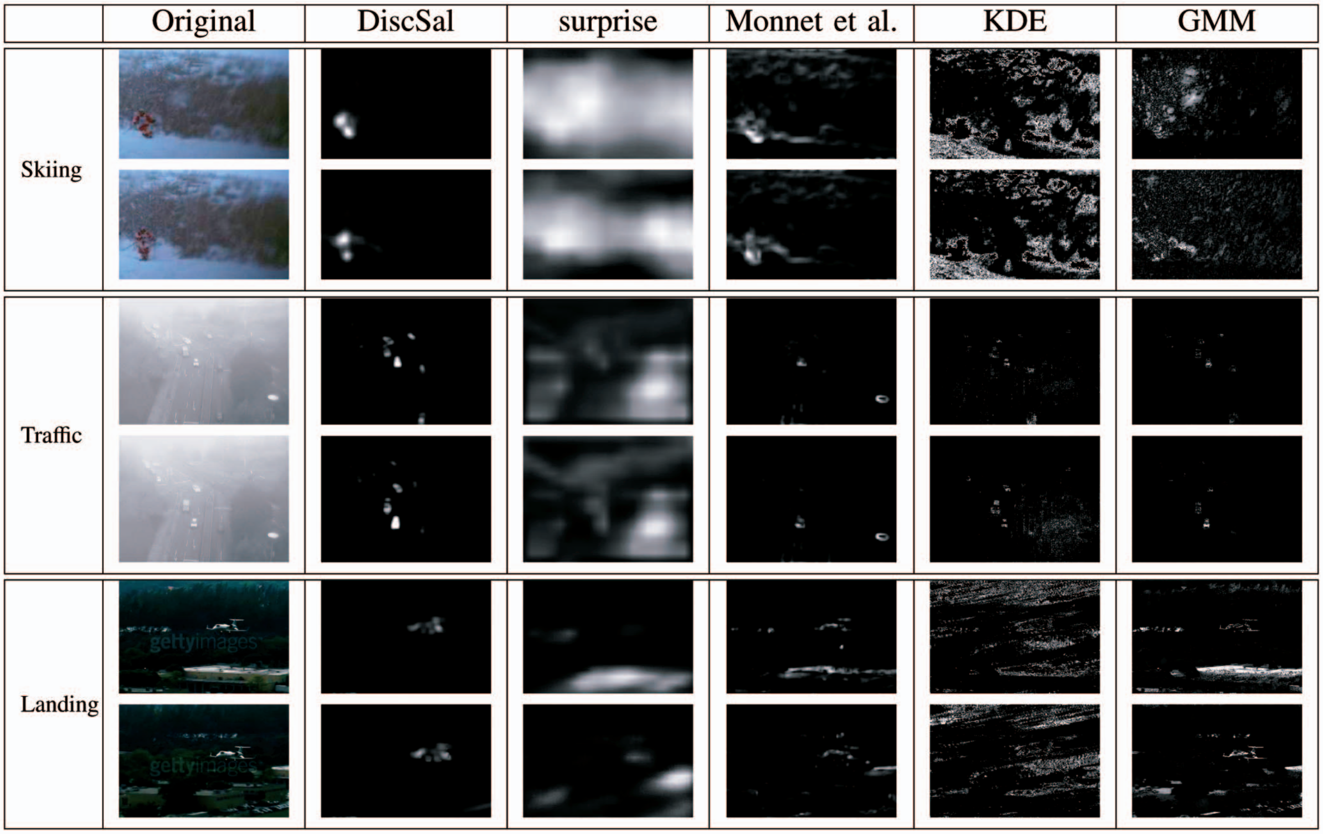


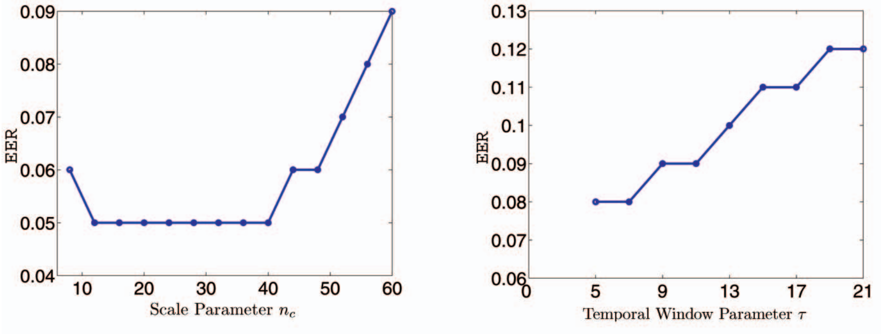
Fig. 3. Saliency maps for three dynamic scenes. For each method, we display a measure of the confidence of each image location being a foreground pixel. In the terminology of the respective publications: DiscSal—discriminant saliency, KDE and GMM—1-probability of background, Monnet—Mahalanobis outlier measure, and Surprise—combined surprise measure.

4.2 Quantitative Analysis

To enable a quantitative analysis, 50-100 frames of each sequence were manually annotated with foreground object segmentation ground truth (a segmentation mask), as perceived by a human observer. All saliency maps were then thresholded at a large number of values, and the false alarm (α) and detection rate (β) computed for each threshold. The resulting receiver operating characteristic (ROC) curves were then summarized by the equal error rate (EER), i.e., the error at which false alarm equals miss rate ($\alpha = 1 - \beta$).

We started by investigating the sensitivity of discriminant saliency to its two free parameters, the size n_c of the center window and the number of frames τ in the temporal window. In the first case, we used a sequence ("birds") whose foreground objects have average size among the sequences considered, set $\tau = 11$, and measured the EER as a function of n_c. In the second, we used a sequence ("boat") with a fast moving foreground object, set $n_c = 16$, and measured EER as a function of τ. The EER curves are presented in Fig. 4, showing that the error rate remains approximately constant over a substantial range of spatiotemporal window sizes. This implies that, while performance improvements would be possible by searching for the spatiotemporal window size that maximizes saliency for each sequence, the additional complexity of this search is usually not warranted. In all subsequent experiments, we have used $n_c = 16$ and $\tau = 11$.

Table 1 shows the EERs of the various methods (DiscSal, DiscSal-Causal, Surprise, KDE, Monnet, and GMM, referred to in the table as DS, DS(C), Su, KDE, Mo, and GMM, respectively) on all sequences, as well as the average over the sequence set. The proposed method outperformed all others, achieving an average



(a) (b)

Fig. 4. Sensitivity of discriminant saliency to (a) the spatial scale parameter n_c on "birds" and (b) the temporal scale parameter τ on "boat."

TABLE 1
Equal Error Rates of Different Background Subtraction Algorithms

	DS	DS(C)	Su	KDE	Mo	GMM
skiing	**3%**	4%	26%	46%	11%	36%
surf	**4%**	5%	30%	36%	10%	23%
cyclists	**8%**	19%	41%	44%	28%	36%
birds	**5%**	9%	19%	20%	7%	23%
chopper	**5%**	**5%**	13%	43%	8%	35%
flock	**15%**	17%	23%	33%	31%	34%
boat	**9%**	11%	**9%**	13%	15%	15%
jump	**15%**	**15%**	25%	51%	23%	39%
surfers	**7%**	8%	24%	25%	10%	35%

	DS	DS(C)	Su	KDE	Mo	GMM
bottle	**2%**	3%	5%	38%	17%	25%
hockey	**24%**	27%	28%	35%	29%	39%
land	**3%**	5%	31%	54%	16%	40%
zodiac	**1%**	**1%**	19%	20%	3%	40%
peds	**7%**	**7%**	37%	17%	11%	11%
traffic	**3%**	4%	46%	39%	9%	34%
freeway	**6%**	10%	43%	21%	31%	25%
ocean	**11%**	**11%**	42%	19%	**11%**	30%
rain	**3%**	6%	10%	23%	17%	15%

	DS	DS(C)	Su	KDE	Mo	GMM
Avg	**7.6%**	9.3%	26.2%	33.1%	16%	29.7%

The average over all sequences is shown in the last row.

EER of 7.6 percent (DiscSal) versus 16 percent for the closest competitor (the method of Monnet et al. [21]). Analyzing the sequences individually, the proposed algorithm exhibits substantial robustness, performing well in the presence of camera motion, variable foreground object scales, and low imaging quality (e.g., scenes with falling snow, fog, and rain). The greatest difficulties occur in scenes, such as "hockey" and "jump," where the foreground objects cover a substantial portion of the image area. In these cases, center surround processing is difficult, due to relative absence of background information. Nevertheless, performance is still superior to those of all previously available methods.

It is also interesting to compare the performance of the different algorithms in light of their saliency representation. There are at least two significant differences between the previous methods and that now proposed. First, the GMM, KDE, and "surprise" models lack a sophisticated unified representation for the spatial and temporal components of saliency. For complex dynamic scenes, where local variation in the background (either spatially or temporally) is significant, this leads to many false positives. The dynamic texture representation is a significant asset in this respect. Second, both the Monnet et al. and GMM/KDE approaches rely uniquely on models of the background, treating foreground objects as outliers. For highly dynamic scenes, it is difficult to account for the large variability of background pixels with a single model. The discriminant nature of the proposed saliency framework is a significant asset in this respect. Overall, both the discriminant formulation and the unified spatiotemporal representation seem to be necessary for good performance. This can be seen from the relative error rates of the various techniques, as shown in Table 1. The algorithm now proposed (DS) exhibits both properties and performs best. Methods that exhibit only one property ("surprise" discriminates between prior and posterior distributions, and Monnet relies on a spatiotemporal representation similar to that of DS) achieve the next best levels of performance. Finally, methods that lack the two properties (GMM and KDE) perform poorly.

5 CONCLUSION

In this work, we proposed an algorithm for spatiotemporal saliency based on a center-surround framework. The new algorithm is inspired by biological vision, namely the psychophysics of motion-based perceptual grouping, and extends a discriminant formulation of center-surround saliency previously proposed for static imagery [13]. This extension is based on the representation of video with dynamic texture models, and is applicable to dynamic scenes. The algorithm combines spatial and temporal components of saliency in a principled manner, and is completely unsupervised. The combination of the discriminant center-surround saliency framework with the modeling power of dynamic textures leads to a robust and versatile procedure for background subtraction, which is successful even for scenes with highly dynamic backgrounds and a moving camera.

The main shortcoming of the current implementation of the proposed algorithm is its computational performance. The processing of each video frame (of size 240×320 pixels) currently requires approximately 37 seconds, for a MATLAB implementation on a PC with 3 GHz CPU and 2 GB RAM. Although we have not so far devoted any attention to computational optimization, we do not expect the algorithm to be deployable in real time without further investigation. We intend to address this issue in future work.

ACKNOWLEDGMENTS

This research was supported by US National Science Foundation Awards IIS-0448609 and IIS-0534985. The authors thank Professor Ahmed Elgammal for providing the code for nonparametric background subtraction using kernel density estimates from [12], Professor Stan Sclaroff for providing some of the test sequences, and Antoni Chan and Dashan Gao for useful discussions.

REFERENCES

[1] http://staff.science.uva.nl/~zivkovic/download.html, 2009.

[2] http://www.svcl.ucsd.edu/projects/background_subtraction, 2009.

[3] M.M. Bence, P. Ölveczky, and S.A. Baccus, "Segregation of Object and Background Motion in the Retina," *Nature*, vol. 423, pp. 401-408, 2003.

[4] R.T. Born, J. Groh, R. Zhao, and S.J. Lukasewycz, "Segregation of Object and Background Motion in Visual Area MT: Effects of Microstimulation on Eye Movements," *Neuron*, vol. 26, pp. 725-734, 2000.

[5] T. Boult, "Coastal Surveillance Datasets," Vision and Security Lab, Univ. of Colorado at Colorado Springs, www.vast.uccs.edu/tboult/PETS2005, 2005.

[6] A. Bugeau and P. Perez, "Detection and Segmentation of Moving Objects in Highly Dynamic Scenes," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[7] A.B. Chan and N. Vasconcelos, "Efficient Computation of the kl Divergence between Dynamic Textures," Technical Report SVCL-TR-2004-02, Dept. of Electrical and Computer Eng., Univ. of California, San Diego, 2004.

[8] A.B. Chan and N. Vasconcelos, "Probabilistic Kernels for the Classification of Auto-Regressive Visual Processes," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 846-851, 2005.

[9] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, 1991.

[10] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting Moving Objects, Ghosts, and Shadows in Video Streams," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1337-1342, Oct. 2003.

[11] G. Doretto, A. Chiuso, Y.N. Wu, and S. Soatto, "Dynamic Textures," *Int'l J. Computer Vision*, vol. 51, no. 2, pp. 91-109, 2003.

[12] A. Elgammal, R. Duraiswami, D. Harwood, and L.S. Davis, "Background and Foreground Modeling Using Nonparametric Kernel Density for Visual Surveillance," *Proc. IEEE*, vol. 90, no. 7, pp. 1151-1163, July 2002.

[13] D. Gao and N. Vasconcelos, "Decision-Theoretic Saliency: Computational Principle, Biological Plausibility, and Implications for Neurophysiology and Psychophysics," *Neural Computation,* vol. 21, pp. 239-271, 2007.

[14] E. Hayman and J. Eklundh, "Statistical Background Subtraction for a Mobile Observer," *Proc. Int'l Conf. Computer Vision,* 2003.

[15] D.H. Hubel and T.N. Wiesel, "Receptive Fields and Functional Architecture in Two Nonstriate Visual Areas (18 and 19) of the Cat," *J. Neurophysiology,* vol. 28, pp. 229-289, 1965.

[16] M. Irani, B. Rousso, and S. Peleg, "Computing Occluding and Transparent Motions," *Int'l J. Computer Vision,* vol. 12, pp. 5-16, 1994.

[17] L. Itti, "The iLab Neuromorphic Vision C++ Toolkit: Free Tools for the Next Generation of Vision Algorithms," *The Neuromorphic Eng.,* vol. 1, no. 1, p. 10, Mar. 2004.

[18] L. Itti and P. Baldi, "A Principled Approach to Detecting Surprising Events in Video," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 631-637, 2005.

[19] L. Itti and C. Koch, "A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention," *Vision Research,* vol. 40, pp. 1489-1506, 2000.

[20] S. Kullback, *Information Theory and Statistics.* Dover Publications, 1968.

[21] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh, "Background Modeling and Subtraction of Dynamic Scenes," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1305-1312, 2003.

[22] H.C. Nothdurft, "The Role of Features in Preattentive Vision: Comparison of Orientation, Motion and Color Cues," *Vision Research,* vol. 33, no. 14, pp. 1937-1958, 1993.

[23] Y. Ren, C. Chua, and Y. Ho, "Motion Detection with Nonstationary Background," *Machine Vision and Applications,* vol. 13, nos. 5-6, pp. 332-343, 2003.

[24] Y. Sheikh and M. Shah, "Bayesian Modeling of Dynamic Scenes for Object Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 11, pp. 1778-1792, Nov. 2005.

[25] C. Stauffer and W. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 2246-2252, 1999.

[26] N. Vasconcelos, "Feature Selection by Maximum Marginal Diversity: Optimality and Implications for Visual Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 762-769, 2003.

[27] R. Visser, N. Sebe, and E. Bakker, "Object Recognition for Video Retrieval," *Proc. Int'l Conf. Image and Video Retrieval,* pp. 250-259, 2002.

[28] L. Wixson, "Detecting Salient Motion by Accumulating Directionally-Consistent Flow," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 8, pp. 774-780, Aug. 2000.

[29] C.R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 780-785, July 1997.

[30] A. Yilmaz, O. Javed, and M. Shah, "Object Tracking: A Survey," *ACM Computing Surveys,* vol. 38, no. 4, p. 13, 2006.

[31] J. Zhong and S. Sclaroff, "Segmenting Foreground Objects from a Dynamic Textured Background via a Robust Kalman Filter," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 1, pp. 44-50, 2003.

[32] Z. Zivkovic, "Improved Adaptive Gaussian Mixture Model for Background Subtraction," *Proc. Int'l Conf. Pattern Recognition,* 2004.