UNSUPERVISED MOVING TARGET DETECTION IN DYNAMIC SCENES

Vijay Mahadevan and Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093

ABSTRACT

We present an unsupervised algorithm for detection of moving targets in highly dynamic scenes. These are scenes whose background is subject to stochastic motion, due to the presence of multiple moving objects (crowds), water, trees swaying in the wind, etc. The algorithm is inspired by biological vision. Target detection is posed as a problem of center-surround *saliency*, which aims to identify the locations of the visual field of maximal contrast with the background. Contrast is defined in terms of both appearance and motion dynamics, and measured using mutual information between stochastic models, known as dynamic textures, which can account for complex motion. This enables very robust target detection in the classes of scenes which have traditionally proven most adverse to tracking. Extensive tests in the context of dynamic background subtraction have shown significantly superior performance to previous techniques.

1. INTRODUCTION

In a natural scene, objects of interest often move amidst complicated backgrounds that are themselves in motion e.g. swaying trees, moving water, waves and rain. The visual system of animals is well adapted to recognizing the most important moving object (referred to henceforth as the "target"), in such scenes. In fact, this ability is central to survival, for instance, by aiding in the identification of potential predators or prey while ignoring unimportant motion in the background. Apart from the obvious importance in visual systems of the biological world, target detection is extremely useful for various computer vision applications such as object recognition in video, activity and gesture recognition, tracking, surveillance and video analysis. For instance, a robot or an autonomous vehicle could benefit from a module to identify objects approaching it amidst possibly moving backgrounds like dust storms, to do effective path planning.

However, unsupervised moving target detection, often posed as the related problem of background subtraction, is hard to solve using conventional techniques in computer vision(see (Sheikh & Shah, 2005) for a review). Extracting the foreground object moving in a scene where the background itself is dynamic is so complex that even though background subtraction is a classic problem in computer vision, there has been relatively little progress for these types of scenes.

A common assumption underlying many techniques for background subtraction is that the camera capturing the scene is static. (Stauffer & Grimson, 1999; Elgammal, Harwood, & Davis, 2000; Wren, Azarbayejani, Darrell, & Pentland, 1997; Monnet, Mittal, Paragios, & Ramesh, 2003; Tavakkoli, Nicolescu, & Bebis, 2006). However, this assumption places severe restrictions on the applicability of such techniques to real-world video clips, that are often shot with hand-held cameras or even on a moving platform in the case of autonomous vehicles. Conventional techniques to address this problem involve explicit camera motion compensation (Jung & Sukhatme, 2004), followed by stationary camera background subtraction techniques. But these methods are cumbersome and require a reliable estimate of the global motion. In extreme cases, when the background itself is highly dynamic, a unique global motion itself may not be possible to estimate.

Another disadvantage of most current approaches is that they model the background explicitly and assume that the algorithm will initially be presented with frames containing only the background (Monnet et al., 2003; Stauffer & Grimson, 1999; Zivkovic, 2004). The background model is built using this data, and regions or pixels that deviate from this model are considered part of the target or foreground. Hence, these techniques are supervised, and the initial phase could be thought of as *training* the algorithm to learn the background parameters. The need to train such algorithms for each scene separately limits their ability to be deployed for automatic surveillance tasks, where manual re-training of the module to operate in each new scene is not feasible.

A further shortcoming in typical algorithms is that they often make unjustified assumptions on the motion characteristics of the target. For instance, it is often assumed that the foreground moves in a consistent direction (temporal persistence) (Wixson, 2000; Li, 2004; Bugeau & Perez, 2007), with more rapid appearance changes than the background (Sheikh & Shah, 2005). However, these are not always valid, espe-

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **01 DEC 2008**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE
Unsupervised Moving Target Detection In Dynamic Scenes

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Electrical and Computer Engineering University of California, San Diego La Jolla, CA 92093

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

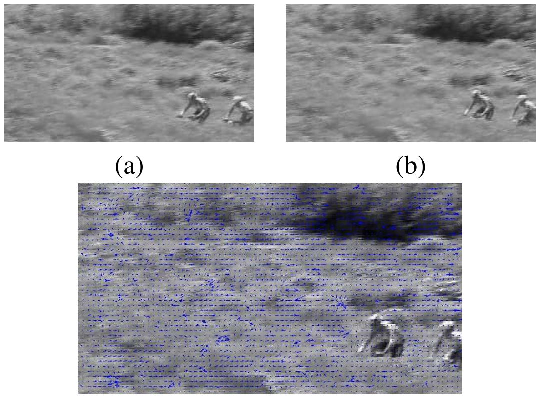
12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM002187. Proceedings of the Army Science Conference (26th) Held in Orlando, Florida on 1-4 December 2008, The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**7**	



(a) (b)

(c)

Fig. 1. (a) and (b) Two consecutive frames from a video clip with camera motion. (c) the optical flow information overlaid on (a). There is no consistent pattern of optical flow in the foreground region in the image.

cially when there is egomotion. As an illustration, two consecutive frames from a video clip shot with a moving camera are shown in Figures 1(a) and (b). The camera panning is such that the objects of interest, viz. the two cyclists, undergo very small motion in the image coordinates. Figure 1(c) shows the optical flow between the two frames. The background is changing rapidly and there is no consistent pattern of flow vectors in the foreground region. The inversion (with respect to the stationary camera scenario) of the motion characteristics of background (which is, in this case, fast moving and temporally coherent) and foreground (whose motion is barely existent and mostly random) can be a major challenge for existing background subtraction techniques.

To address these limitations of existing algorithms, we propose a novel paradigm for unsupervised target detection using motion saliency. The algorithm is based on the idea that in the absence of high-level goals (such as explicit search for a known object) the target consists of the *most salient locations* of the visual field. Salient locations in turn are those that enable the discrimination between center and surround at that location with smallest expected probability of error. This is formalized in a biologically inspired framework referred to as the *discriminant center-surround hypothesis* (Gao & Vasconcelos, 2005, 2007) and, by definition, produces saliency measures that are optimal in a classification sense. This framework can be applied to any type of stimuli and features, and optimal saliency detectors have already been derived for various stimulus modalities for static images, including color and orientation (Gao & Vasconcelos, 2007). In this work, we extend the notion of discriminant center-surround saliency to moving stimuli. By defining saliency in a discriminant sense, we eliminate the need to separately model the background or the target. A single model for representing the motion of a region of the video is sufficient and the most salient moving object is simply the one that best stands out among other objects in the video with respect to this model. As the algorithm compares the regions against one another, it depends only on the *relative disparity* between their motion characteristics, and therefore is invariant to camera motion.

In order to extend this architecture to moving stimuli, probabilistic models that capture the motion patterns in video are needed. In this work, we choose dynamic textures (Doretto, Chiuso, Wu, & Soatto, 2003) to model motion due to their versatility in modeling complex moving patterns and the rich statistical formulations they lend themselves to. In particular, dynamic textures provide a unified generative stochastic model for appearance as well as motion, and these can be conveniently incorporated into a discriminant center-surround framework.

The main contributions of this work are as follows. (a) The proposed algorithm is completely unsupervised and does not require initial training. This enables the algorithm to adapt to any scene without manual intervention. (b) By modeling the video sequences using dynamic textures, saliency in motion and appearance are both taken into account in a principled manner, without the need to model either explicitly. The proposed discriminant motion saliency algorithm can automatically distinguish between object and background motion due to the distinct appearance and motion characteristics of the two regions. (c) Finally, being a discriminant technique, the algorithm ignores egomotion, and can handle video clips shot with a moving camera.

The remaining sections of the paper are organized as follows: the discriminant saliency architecture is presented in Section 2. Representation of the target and background using dynamic texture models are discussed in Section 3. The target detection algorithm is summarized in 4. Experimental evaluation and results form Section 5.

2. DISCRIMINANT CENTER-SURROUND SALIENCY

Discriminant saliency (Gao & Vasconcelos, 2007) is defined with respect to two classes of stimuli: the class of *stimuli of interest*, and the *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 which opposes the stimuli of interest to the null hypothesis. The saliency of each location in the visual field is then equated to the discriminant power (expected classification accuracy) of the visual features extracted from that location in differentiating the two classes.

Formally, let $\mathcal{V}$ be a d dimensional dataset indexed by location vector $l \in L \subset \mathbb{R}^d$ and consider the responses to visual stimuli of a predefined set of features $\boldsymbol{Y}$ (e.g. raw pixel values, Gabor or Fourier features), computed from $\mathcal{V}$ at all locations

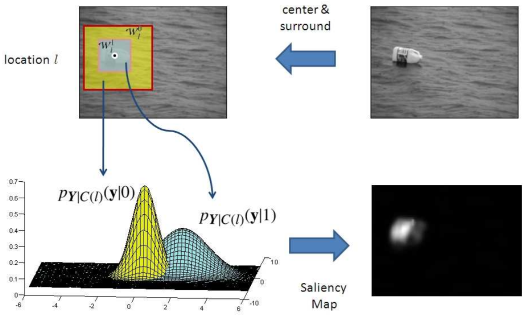


Fig. 2. Illustration of discriminant center-surround saliency. Center and surround windows are defined around each image location, and the distribution of a previously defined set of features $\boldsymbol{Y}$ estimated from the two windows. The saliency of the location is a measure of how disjoint the two feature distributions are.

$l \in L$. 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 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 $p_{Y|C(l)}(\mathbf{y}|1)$. 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 $p_{Y|C(l)}(\mathbf{y}|0)$. An illustration of the center-surround classification problem, for a static image, is shown in Figure 2.

The saliency of location l, $S(l)$, is the extent to which the features $\boldsymbol{Y}$ can discriminate between *center* and *surround*. This is quantified by the mutual information between features, $\boldsymbol{Y}$, and class label, C,

$$\begin{aligned} S(l) &= I_l(\boldsymbol{Y}; C) \\ &= \sum_{c=0}^{1} \int p_{Y,C(l)}(\mathbf{y}, c) \log \frac{p_{Y,C(l)}(\mathbf{y}, c)}{p_Y(\mathbf{y}) p_{C(l)}(c)} d\mathbf{y}. \end{aligned} \quad (1)$$

This mutual information is an approximation to the expected probability of correct classification (more precisely one minus the Bayes error rate) of the classification problem that opposes center to surround (Vasconcelos, 2003). So, a large value of saliency $S(l)$ implies that center and surround have a large *local feature contrast*, which enables their discrimination with low probability of error. Conversely, the locations where the classification as a target has the smallest expected probability of error can be identified by searching for maxima of $S(l)$. The function $S(l), l \in L$ is referred to as the *saliency map* of the dataset $\mathcal{V}$. It can also be written as

$$\begin{aligned} S(l) &= \sum_{c=0}^{1} p_{C(l)}(c) \int p_{Y|C(l)}(\mathbf{y}|c) \log \frac{p_{Y|C(l)}(\mathbf{y}|c)}{p_Y(\mathbf{y})} d\mathbf{y} \quad (2) \\ &= \sum_{c=0}^{1} p_{C(l)}(c) \mathrm{KL}\left(p_{Y|C(l)}(\mathbf{y}|c) \,||\, p_Y(\mathbf{y})\right) \quad (3) \end{aligned}$$

where

$$\mathrm{KL}\left(p \,||\, q\right) = \int_y p_Y(y) \log \frac{p_Y(y)}{q_Y(y)} dy.$$

is the Kullback-Leibler (KL) divergence between the probability distributions $p_X(x)$ and $q_X(x)$ (Kullback, 1968). This allows an alternative interpretation of saliency as a measure of the average distance between the feature distribution over each window and the average of the two distributions. This is a measure of the (lack of) overlap between the distributions associated with center and surround.

3. REPRESENTATION OF VIDEO USING DYNAMIC TEXTURES

The discriminant saliency formulation of (1) is generic and does not vary with the type of stimulus or features $\boldsymbol{Y}$ used.

In specific, by adopting suitable models for spatiotemporal stimuli (i.e. video), this formulation is robust enough to compute motion saliency in highly dynamic scenes. This enables the design of powerful target detection algorithms by simple reduction of target detection to the complement of saliency detection. Under this formulation, the design of an algorithm capable of handling highly dynamic scenes only requires the use, in (3), of probability models $p_{Y|C(l)}(\mathbf{y}|c)$ that can capture the variability associated with such video scenes. We adopt the dynamic texture (DT) model of (Doretto et al., 2003), due to its ability to account for this variability, while jointly modeling the spatial and temporal characteristics of the visual stimulus in an elegant unified stochastic framework.

A dynamic texture is an autoregressive generative model that represents the appearance of the stimulus $\mathbf{y}_t \in \mathbb{R}^m$ (the two-dimensional image stimulus is first converted into a column vector of length m), observed at time t, as a linear function of a hidden state process $\mathbf{x}_t \in \mathbb{R}^n$ ($n << m$) subject to Gaussian observation noise. The state and appearance processes form a linear dynamical system (LDS)

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{v}_t \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \mathbf{w}_t \end{aligned} \quad (4)$$

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

$$\boldsymbol{\Theta} = (\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}, \boldsymbol{\mu}_1, \mathbf{S}_1). \quad (5)$$

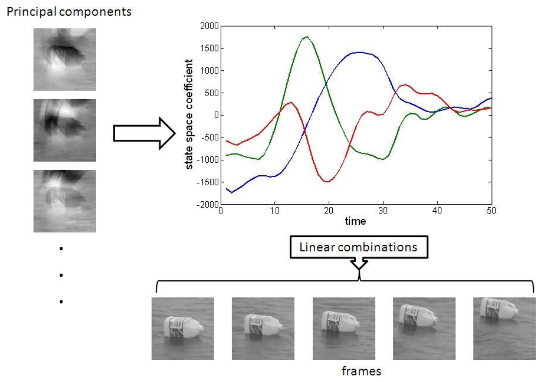


Fig. 3. Illustration of a dynamic texture model. The first three basis images are shown on the left, and the corresponding state space variables plotted as a function of time. At each time instant, a video frame is represented as a linear combination of the basis images, with weights given by the value of the corresponding state variable.

The hidden state space sequence $\mathbf{x}_t$ is a first order Markov chain that encodes stimulus dynamics, while $\mathbf{y}_t$ is a linear combination of prototypical basis functions (the columns of $\mathbf{C}$) and encodes the appearance component of the stimulus at time t. Dynamic texture modeling of a sequence of images is illustrated in Figure 3[1].

3.1. Learning dynamic texture parameters

Given center and surround regions, DT parameters could in principle be learned by maximum likelihood (using expectation-maximization (Shumway & Stoffer, 1982), or N4SID (Overschee & Moor, 1994)). However, due to the high dimensionality of video sequences, these solutions are too complex for motion saliency. A suboptimal alternative, that works well in practice(Doretto et al., 2003), is to learn the spatial and temporal parameters separately. Given N sequences, $\mathbf{y}^{(1)}_{1:\tau}, \ldots, \mathbf{y}^{(N)}_{1:\tau}$, of τ frames each (where $\mathbf{y}^{(i)}_{1:\tau} = [\mathbf{y}^{(i)}_1 \ldots \mathbf{y}^{(i)}_\tau]$), sampled from a DT, let $\mathbf{Y}_{1:\tau} = [\mathbf{y}^{(1)}_{1:\tau}, \ldots, \mathbf{y}^{(N)}_{1:\tau}] \in \mathbb{R}^{m \times N\tau}$ be the matrix composed by concatenating all sequences. If $\mathbf{Y}_{1:\tau} = \mathbf{USV}^T$ is its singular value decomposition (SVD), the DT parameters are estimated as follows,

$$\begin{aligned}
\hat{\mathbf{C}} &= \mathbf{U}[1:n] \text{ (first n columns of } \mathbf{U}) && (6)\\
\hat{\mathbf{x}}^{(i)}_{1:\tau} &= \hat{\mathbf{C}}^T \mathbf{y}^{(i)}_{1:\tau} && (7)\\
\hat{\mathbf{A}} &= \hat{\mathbf{X}}_{2:\tau}(\hat{\mathbf{X}}_{1:\tau-1})^{\dagger} && (8)\\
\hat{\mathbf{Q}} &= \frac{1}{N(\tau-1)} \sum_{i=1}^{N} \sum_{j=1}^{\tau-1} \hat{\mathbf{v}}^{(i)}_j (\hat{\mathbf{v}}^{(i)}_j)^T && (9)
\end{aligned}$$

$$\begin{aligned}
\hat{\mathbf{R}} &= \frac{1}{N(\tau-1)} \sum_{i=1}^{N} \sum_{j=1}^{\tau-1} \hat{\mathbf{w}}^{(i)}_j (\hat{\mathbf{w}}^{(i)}_j)^T && (10)\\
& && (11)
\end{aligned}$$

where, $\hat{\mathbf{X}}_{1:\tau} = [\hat{\mathbf{x}}^{(1)}_{1:\tau}, \ldots, \hat{\mathbf{x}}^{(N)}_{1:\tau}]$ is the matrix of state estimates, $\mathbf{M}^{\dagger}$ the pseudo-inverse of $\mathbf{M}$, $\hat{\mathbf{v}}^{(i)}_t = \hat{\mathbf{x}}^{(i)}_{t+1} - \hat{\mathbf{A}}\hat{\mathbf{x}}^{(i)}_t$, and $\hat{\mathbf{w}}^{(i)}_t = \mathbf{y}^{(i)}_t - \hat{\mathbf{C}}\hat{\mathbf{x}}^{(i)}_t$, for $t \in 1 \ldots \tau$. Finally, the initial state parameters are estimated as,

$$\begin{aligned}
\hat{\mu}_1 &= \frac{1}{N} \sum_{i=1}^{N} \hat{\mathbf{x}}^{(i)}_1 && (12)\\
\hat{\mathbf{S}}_1 &= \frac{1}{N} \sum_{i=1}^{N} \hat{\mathbf{x}}^{(i)}_1 (\hat{\mathbf{x}}^{(i)}_1)^T - \hat{\mu}_1 \hat{\mu}_1{}^T && (13)
\end{aligned}$$

Using the learned model parameters, we can compute probability distributions over the DT. 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). So for Gaussian initial state conditions, the joint distribution of the state sequence, $\mathbf{x}_{1:\tau} = [\mathbf{x}^T_1 \ldots \mathbf{x}^T_\tau]^T$ is also Gaussian (Chan & Vasconcelos, 2005)

$$p(\mathbf{x}_{1:\tau}) = G(\mathbf{x}_{1:\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad (14)$$

where $\boldsymbol{\mu} = \begin{bmatrix} \mu^T_1 & \cdots & \mu^T_\tau \end{bmatrix}^T$ and the covariance is

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{S}_1 & (\mathbf{AS}_1)^T & \cdots & (\mathbf{A}^{\tau-1}\mathbf{S}_1)^T \\ \mathbf{AS}_1 & \mathbf{S}_2 & \cdots & (\mathbf{A}^{\tau-2}\mathbf{S}_2)^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}^{\tau-1}\mathbf{S}_1 & \mathbf{A}^{\tau-2}\mathbf{S}_2 & \cdots & \mathbf{S}_\tau \end{bmatrix}. \qquad (15)$$

Similarly, the image sequence $\mathbf{y}_{1:\tau}$ is distributed as

$$p(\mathbf{y}_{1:\tau}) = G(\mathbf{y}_{1:\tau}, \boldsymbol{\gamma}, \boldsymbol{\Phi}) \qquad (16)$$

where $\boldsymbol{\gamma} = \boldsymbol{C}\boldsymbol{\mu}$ and $\boldsymbol{\Phi} = \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^T + \boldsymbol{\mathcal{R}}$, and $\boldsymbol{C}$ and $\boldsymbol{\mathcal{R}}$ are block diagonal matrices formed from $\mathbf{C}$ and $\mathbf{R}$ respectively:

$$\boldsymbol{C} = \begin{bmatrix} \mathbf{C} & 0 & \cdots & 0 \\ 0 & \mathbf{C} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{C} \end{bmatrix}, \boldsymbol{\mathcal{R}} = \begin{bmatrix} \mathbf{R} & 0 & \cdots & 0 \\ 0 & \mathbf{R} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{R} \end{bmatrix}.$$

At any location l of the video, the densities of (15) can be estimated from a collection of spatio-temporal patches extracted from the center and surround windows. The evaluation of the KL divergence between DTs is needed for the computation of $S(l)$,with (3).

Let $p_0(\mathbf{y}_{1:\tau})$ and $p_1(\mathbf{y}_{1:\tau})$ be the probabilities of a sequence of τ frames under two DTs parameterized by $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$, corresponding to the surround and the center respectively. For

[1] The bottle sequence from (Zhong & Sclaroff, 2003) is used in this example.

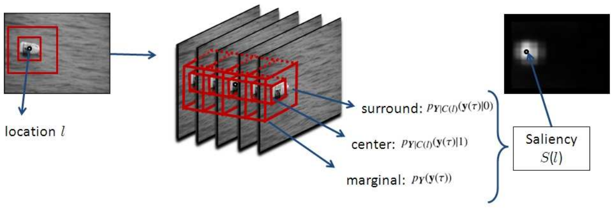

location l
surround: $p_{Y|C(l)}(\mathbf{y}(\tau)|0)$
center: $p_{Y|C(l)}(\mathbf{y}(\tau)|1)$
marginal: $p_Y(\mathbf{y}(\tau))$
Saliency
$S(l)$

Fig. 4. 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 measure $S(l)$ is finally computed with (3).

Gaussian p_0 and p_1, the KL divergence has the closed-form (Cover & Thomas, 1991):

$$\begin{aligned} &\mathrm{KL}\left(p_0 \| p_1\right) \qquad (17) \\ &= \frac{1}{2}\left[\log\frac{|\mathbf{\Phi}_1|}{|\mathbf{\Phi}_0|} + \mathrm{tr}\left(\mathbf{\Phi}_1^{-1}\mathbf{\Phi}_0\right) + \left\|\gamma_0 - \gamma_1\right\|^2_{\mathbf{\Phi}_1} - m\tau\right] \end{aligned}$$

where m is the number of pixels in each frame. Direct evaluation of the KL is computationally intractable, since the expression depends on $\mathbf{\Phi}_0$ and $\mathbf{\Phi}_1$, which are very large covariance matrices. An efficient recursive procedure is, however, available (Chan & Vasconcelos, 2004).

4. TARGET DETECTION ALGORITHM

In an unsupervised task, in the absence of any information regarding a specific previously known target, motion saliency provides an objective way of defining the target : the target consists of those regions of the video with high motion saliency. So, moving target detection is performed by first computing the saliency measure $S(l)$ at each location l of the video.

Center and surround windows are centered at the location, and a collection of spatio-temporal patches extracted from each window. Prior probabilities for both classes are assumed to be equal to $\frac{1}{2}$. DT parameters are then learned from the center, surround, and total windows, to obtain the densities $p_{Y|C(l)}(\mathbf{y}(\tau)|1)$, $p_{Y|C(l)}(\mathbf{y}(\tau)|0)$, and $p_Y(\mathbf{y}(\tau))$, respectively. $S(l)$ is finally computed with (3), a the recursive implementation of (16) (Chan & Vasconcelos, 2004). The procedure is illustrated in Figure 4. Those pixels which have a saliency value above a predetermined threshold are marked as belonging to the moving target. The motion saliency based target detection algorithm is summarized in Algorithm 1.

Algorithm 1 Target Detection via Computation of 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 $\mathcal{W}_l^1$ and surround $\mathcal{W}_l^0$.
    List all overlapping patches of size $n_p \times n_p \times \tau$ in $\mathcal{W}_l^1$ and $\mathcal{W}_l^0$
    From the patches learn dynamic texture parameters for center $\mathbf{\Theta}_1(l)$, surround $\mathbf{\Theta}_0(l)$ and the total $\mathbf{\Theta}(l)$ using (5)-(12).
    Compute the mutual information, $S(l)$, between class-conditional and total densities (3), using the recursive implementation of (16).
**end for**
Choose threshold value $T$. Find regions where $l_{target} = \{l \in L : S(l) > T\}$.
**Output:** Target locations $l_{target}$
```

5. EXPERIMENTS AND RESULTS

To evaluate target detection performance, the proposed algorithm was tested on sequences collected from the web. Frames from some of these sequences are shown in panel (a) of Figures 5 - 7. In all cases, the background is highly dynamic. In addition, most sequences were shot with significant camera motion. Figure 5, presents frames from a sequence which depicts a person skiing in heavy snowfall. A pair of cyclists ride through a grassy plain in Figure 6, while an aircraft landing is tracked using a moving camera in Figure 7. Due to the extreme variability in background these clips are challenging for conventional foreground detection techniques.

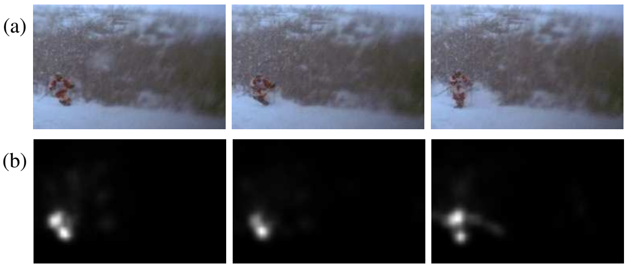


Fig. 5. Results of target detection on a skiing clip shot with a moving camera, with heavy snowfall in the background: (a) original (b) detected target

To perform target detection, the sequences were converted to grayscale, and saliency maps computed at sub-sampled locations of the video, using a grid scaled down by a factor of 4 spatially and 2 temporally. At each grid location, the center window occupied 16 × 16 pixels and spanned 11 frames - 5 past frames, the current frame, and 5 frames in the future.

Saliency maps obtained using the proposed algorithm on the test clips are shown in panel (b) of Figures 5, 6 and 7. Video sequences of these and various other detection examples are available from `www.svcl.ucsd.edu/~projects/background_subtraction`. Even though the background is extremely dynamic, the relevant targets are detected accu-

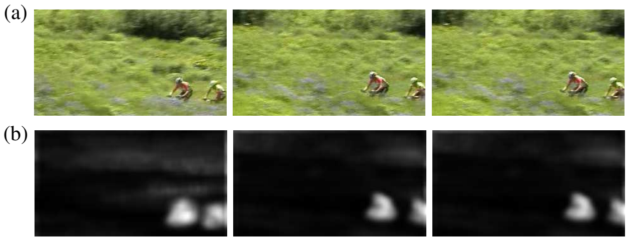


Fig. 6. Results of target detection on clip showing a pair of cyclists. The camera is moving to track the cyclists, causing very large variability in the background: (a) original (b) detected targets

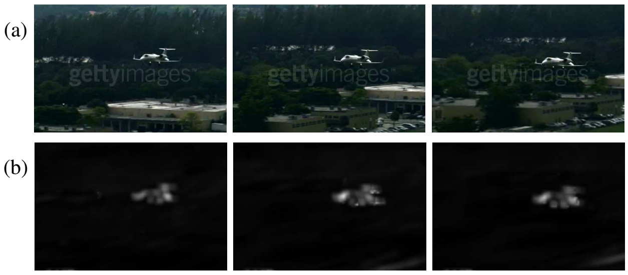


Fig. 7. Results of target detection on clip showing an aircraft landing. The camera is moving to keep the aircraft in focus, causing variability in the background which consists of buildings, cars and trees: (a) original (b) detected target

rately, in all three cases.

To enable a quantitative analysis, all sequences were manually annotated with the groundtruth for the objects of interest. The saliency maps were then thresholded at a large number of values, and using the groundtruth information false alarm (α) and detection rate (β) were computed. These were used to generate receiver operating characteristic (ROC) curves. Using the ROC curves, the equal error rate (EER), defined as the error at which false alarm equals miss rate ($\alpha = 1-\beta$), was also estimated. The EER represents a quantitative measure of target detection performance of the proposed algorithm. The low EER (average of around 4.7% shows that the proposed algorithm identifies the target reliably with low false positive rate. Table 1 shows the EERs for the three clips of Figures 5 - 7.

	EER
skiing	**3%**
cyclists	**8%**
landing	**3%**
Average	**4.7%**

Table 1. Equal Error Rates for the sequences tested. The proposed algorithm has very low EER for all clips, showing that it can accurately detect the target with very low false postive rate.

6. CONCLUSION

In this work, we have proposed an algorithm for unsupervised moving target detection based on center-surround saliency. The new algorithm is inspired by biological vision, and extends a discriminant formulation of center-surround saliency previously proposed for static imagery. By using dynamic texture models for motion, we derive an information theoretic measure of motion saliency. The discriminant center-surround framework, in combination with the modeling power of dynamic textures leads to a robust and versatile algorithm that can be applied to scenes with highly dynamic backgrounds, even when the camera is moving. The algorithm combines spatial and temporal components of saliency in a principled manner. Being completely unsupervised it does not require any training and can thus be automatically deployed to new scenes, with no need for manual supervision or parameter tuning. As the algorithm can work even for moving cameras, it can also be incorporated into hand-held or vehicle mounted sensing devices.

Potential applications for the army include automated surveillance with alerts for specific events, detection of events in archived video, crowd monitoring, detection of breaches of borders and other secure areas, path planning for autonomous vehicles and automated target tracking.

ACKNOWLEDGMENT

This research was supported by NSF awards IIS-0448609, and IIS-0534985.

References

Bugeau, A., & Perez, P. (2007). Detection and segmentation of moving objects in highly dynamic scenes. In *Computer vision and pattern recognition.*

Chan, A. B., & Vasconcelos, N. (2004). *Efficient computation of the kl divergence between dynamic textures* (Tech. Rep. No. SVCL-TR-2004-02). Dept. of ECE, UCSD.

Chan, A. B., & Vasconcelos, N. (2005). Probabilistic kernels for the classification of auto-regressive visual processes. In *Computer vision and pattern recognition* (Vol. 1, pp. 846–851).

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: John Wiley & Sons Inc.

Doretto, G., Chiuso, A., Wu, Y. N., & Soatto, S. (2003). Dynamic textures. *International Journal of Computer Vision*, *51*(2), 91-109.

Elgammal, A., Harwood, D., & Davis, L. (2000). Nonparametric model for background subtraction. In *European conference on computer vision* (p. 751-757).

Gao, D., & Vasconcelos, N. (2005). Discriminant saliency for visual recognition from cluttered scenes. In *Pro-*

ceedings neural information processing systems. Vancouver, Canada.

Gao, D., & Vasconcelos, N. (2007). Decision-theoretic saliency: computational principle, biological plausibility, and implications for neurophysiology and psychophysics. submitted to *Neural Computation*.

Jung, B., & Sukhatme, G. S. (2004, March). Detecting moving objects using a single camera on a mobile robot in an outdoor environment. In *International conference on intelligent autonomous systems, amsterdam, the netherlands* (p. 980-987).

Kullback, S. (1968). *Information theory and statistics*. Dover Publications, New York.

Li, Y. (2004). On incremental and robust subspace learning. *Pattern Recognition*, *37*(7), 1509-19.

Monnet, A., Mittal, A., Paragios, N., & Ramesh, V. (2003). Background modeling and subtraction of dynamic scenes. In *Computer vision and pattern recognition.*

Murray, A., D. Basu. (1994). Motion tracking with an active camera. *IEEE Transactions Pattern Analysis and Machine Intelligence*, *16*(5), 449-459.

Overschee, P. V., & Moor, B. D. (1994). N4sid: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, *30*, 75-93.

Sheikh, Y., & Shah, M. (2005). Bayesian modeling of dynamic scenes for object detection. *IEEE Pattern Analysis and Machine Intelligence*, *27*(11), 1778-92.

Shumway, R., & Stoffer, D. (1982). An approach to time series smoothing andforecasting using the EM algorithm. *Journal of Time Series Analysis*, *3*(4), 433–467.

Stauffer, C., & Grimson, W. (1999). Adaptive background mixture models for real-time tracking. In *Ieee computer vision and pattern recognition* (Vol. 2, p. 2246-2252).

Tavakkoli, A., Nicolescu, M., & Bebis, G. (2006). A novelty detection approach for foreground region detection in videos with quasi-stationary backgrounds. In *International symposium on visual computing.*

Vasconcelos, N. (2003). Feature selection by maximum marginal diversity: optimality and implications for visual recognition. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 1, p. 762-769).

Wixson, L. (2000). Detecting salient motion by accumulating directionally-consistent flow. *IEEE Pattern Analysis and Machine Intelligence*, *22*(8), 774-780.

Wren, C. R., Azarbayejani, A., Darrell, T., & Pentland, A. (1997). Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 780-785.

Zhong, J., & Sclaroff, S. (2003). Segmenting foreground objects from a dynamic textured background via a robust Kalman filter. In *Proceedings of IEEE International Conference on Computer Vision* (Vol. 1, p. 44).

Zivkovic, Z. (2004). Improved adaptive Gaussian mixture model for background subtraction. In *International conference on pattern recognition.*