Motion-Based Background Subtraction using Adaptive Kernel Density Estimation

Anurag Mittal
anurag@scr.siemens.com
Real-Time Vision and Modeling
Siemens Corporate Research
Princeton, NJ 08540

Nikos Paragios *
nikos.paragios@computer.org
C.E.R.T.I.S.
Ecole Nationale de Ponts et Chaussees
Champs sur Marne, France

Abstract

Background modeling is an important component of many vision systems. Existing work in the area has mostly addressed scenes that consist of static or quasi-static structures. When the scene exhibits a persistent dynamic behavior in time, such an assumption is violated and detection performance deteriorates. In this paper, we propose a new method for the modeling and subtraction of such scenes. Towards the modeling of the dynamic characteristics, optical flow is computed and utilized as a feature in a higher dimensional space. Inherent ambiguities in the computation of features are addressed by using a data-dependent bandwidth for density estimation using kernels. Extensive experiments demonstrate the utility and performance of the proposed approach.

1 Introduction

Increased computational speed of processors has enabled application of vision technology in several fields such as: Industrial automation, Video security, transportation and automotive. Background subtraction forms an important component in many of these applications. The central idea behind this module is to utilize the visual properties of the scene for building an appropriate representation that can then be utilized for the classification of a new observation as foreground or background. The information provided by such a module can then be considered as a valuable low-level visual cue to perform high-level object analysis tasks such as object detection, tracking, classification and event analysis.

Existing methods for background modeling may be classified as either *predictive* or *non-predictive*. *Predictive* methods model the scene as a time series and develop a dynamical model to recover the current input based on past observations. The magnitude of the deviation between the predicted and actual observation can then be used as a measure of change. Predictive mechanisms of varying complexity have been considered in the literature. Several authors [16, 17] have used a Kalman-filter based approach for modeling the dynamics of the state at a particular pixel. A simpler version of the Kalman filter called *Weiner filter* is considered in [23] that operates directly on the data. Such modeling may further be performed in an appropriate subspace [26, 21] (PCA basis is the usual choice). Recent methods are based on more complicated models. In [7], an autoregressive model was proposed to capture the properties of dynamic scenes. This method was modified in [20, 26] to address the modeling of dynamic backgrounds and perform foreground detection.

The second class of methods (which we call *non-predictive density-based* methods) neglect the order of the input observations and build a probabilistic representation (*p.d.f.*) of the observations at a particular pixel. In [25], a single Gaussian is considered to model the statistical distribution of a background pixel. Friedman et. al.[9] use a mixture of three Normal distributions to model the visual properties in traffic surveillance applications. Three hypothesis are considered - road, shadow and vehicles. The EM algorithm is used, which although optimal, is computationally quite expensive. In [12], this idea is extended by using multiple Gaussians to model the scene and develop a fast approximate method for updating the parameters of the model incrementally. Such an approach is capable of dealing with multiple hypothesis for the background and can be useful in scenes such as waving trees, beaches, escalators, rain or snow. The mixture-of-Gaussians method is quite popular and was to be the basis for a large number of related techniques [15, 13]. In [10], a statistical characterization of the error associated with this algorithm is studied. When the density function is more complex and cannot be modeled

*The work was carried out during the appointment of the author with Siemens Corporate Research, from November 1999 to March 2004

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)
1063-6919/04 $20.00 © 2004 **IEEE**

parametrically, a non-parametric approach able to handle arbitrary densities is more suitable. Such an approach was used in [8] where the use of Gaussian kernels for modeling the density at a particular pixel was proposed.

Existing methods can effectively describe scenes that have a smooth behavior and limited variation. Consequently, they are able to cope with gradually evolving scenes. However, one can claim that their performance deteriorates [Figure 2] when the scene to be described is dynamic and exhibits non-stationary properties in time. Examples of such scenes include ocean waves, waving trees, rain, moving clouds, etc. One can observe that even recent predictive methods do not model multiple modalities of dynamic behavior [20, 26, 23], and therefore such models can be quite inefficient in many practical scenarios.

Most of the dynamic scenes exhibit persistent motion characteristics. Therefore, a natural approach to model their behavior is via optical flow. Combining such flow information with standard intensity information, we present a method for background-foreground differentiation that is able to detect objects that differ from the background in either motion or intensity properties.

Computation of these features, however, has some inherent ambiguities associated with them. The motion of a moving one-dimensional pattern viewed through a circular aperture causes the so-called *aperture problem*. Furthermore, in locations where the spatial gradient vanishes, the motion cannot be estimated. This is sometimes called the *black wall problem*. Transformation of intensity into an illumination-invariant space causes further ambiguities. For very dark regions, such transformation is ill-defined and subject to significant errors.

It is natural to assume that the utilization of such ambiguities in the estimation process will significantly enhance the accuracy of such estimation. To this end, we propose the use of variable bandwidths for density estimation using kernels. Use of such technique not only enables utilization of such ambiguities but also enables modeling of arbitrary shapes of the underlying density in a natural way. Density estimation is performed in a higher-dimensional space consisting of intensity and motion features for the purpose of modeling the background density, and thus perform foreground detection.

The paper is organized as follows. Section 2 describes density estimation via variable-bandwidth kernels. Such an estimation utilizes the uncertainty in both the sample and test observations. Section 3 describes the development of appropriate measures for classification. Section 4 describes methods for computation of optical flow and illumination-invariant intensity transformation. Finally, section 5 describes experiments that quantify the performance of the proposed approach in relation to existing methods for some real-world scenes.

2 Background Density Estimation via Variable Bandwidth Kernels

In order to facilitate the introduction of the proposed framework, we assume that flow measurements [18, 14] and their uncertainties are available. Then, we propose a theoretical framework to obtain an estimate of the probability distribution of the observed data in a higher-dimensional space [1]. Several methods - parametric and non-parametric - can be considered for determining this probability distribution. A mixture of multi-variate Gaussians can be considered to approximate this distribution. The parameters of the model, i.e. the mean and the covariance matrix of the Gaussians, can be estimated and updated in a manner similar to [12]. Care has to be exercised, however, in dealing with the uncertainties in the correct manner.

A more suitable approach refers to a non-parametric method. One can claim that such method has the characteristic of being able to deal with the uncertainties in an accurate manner. On the other hand, such a method is computationally expensive [2].

The most attractive method used in the statistical literature for modeling multi-variate probability distributions from sample points is the *kernel*-based density estimation [24]. (also called *Parzen windows* in Pattern Recognition). Such a selection is even more appropriate when the sample points have variable uncertainties associated with them since the framework provides a structured way for utilizing such uncertainties.

Let $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ be a set of d-dimensional points in $\mathbb{R}^d$ and $\mathbf{H}$ be a symmetric positive definite $d \times d$ matrix (called the *bandwidth* matrix). Let $K : \mathbb{R}^d \rightarrow \mathbb{R}^1$ be a kernel satisfying certain conditions that will be defined later.

Then the multivariate *fixed* bandwidth kernel estimator is defined as[24]:

$$\begin{aligned} \hat{f}(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\|\mathbf{H}\|^{1/2}} K(\mathbf{H}^{-1/2}(\mathbf{x}-\mathbf{x}_i)) \end{aligned} \tag{1}$$

where $K_{\mathbf{H}}(\mathbf{x}) = \|\mathbf{H}\|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$. The matrix $\mathbf{H}$ is the smoothness parameter and specifies the "width" of the kernel around each sample point $\mathbf{x}_i$.

A well-behaved kernel K must satisfy the following con-

[1] A five-dimensional space has been utilized in this work - two components for the optical flow and three for the intensity in the normalized color space.

[2] With the rapid increase in the computational power of processors, this method is already running in quasi real-time (7 fps) on a 160×120 3-band video on a Pentium IV 3 GHz processor machine.

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)
1063-6919/04 $20.00 © 2004 **IEEE**

ditions:

$$\int_{\mathbb{R}^d} K(\mathbf{w})d\mathbf{w} = \mathbf{1},$$
$$\int_{\mathbb{R}^d} \mathbf{w}K(\mathbf{w})d\mathbf{w} = \mathbf{0},$$
$$\int_{\mathbb{R}^d} \mathbf{w}\mathbf{w}^T K(\mathbf{w})d\mathbf{w} = \mathbf{I}_d$$

The first condition accounts for the fact that the sum of the kernel function over the whole region is unity. The second equation imposes the constraint that the means of the *marginal kernels* $\{K_i(w_i), i = 1, \ldots, d\}$ are all zero. Last but not the least, the third term states that the marginal kernels are all pairwise uncorrelated and that each has unit variance.

The simplest approach would be to use a fixed bandwidth matrix $\mathbf{H}$ for all the samples. Although such an approach is a reasonable compromise between complexity and the quality of approximation, the use of variable bandwidth can usually lead to an improvement in the accuracy of the estimated density. Smaller bandwidth is more appropriate in regions of high density since a larger number of samples enables a more accurate estimation of the density in these regions. On the other hand, a larger bandwidth is more appropriate in low density areas where few sample points are available.

It is possible to consider a bandwidth function that adapts to the point of estimation, as well as to the observed data points and the shape of the underlying density[24]. In the literature, two simplified versions have been studied. The first varies the bandwidth at each estimation point and is referred to as the *balloon estimator*. The second varies the bandwidth for each data point and is referred to as the *sample-point estimator*.

Thus, for the *balloon estimator*,

$$\hat{f}_B(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}(\mathbf{x})}(\mathbf{x} - \mathbf{x}_i))$$
$$= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\|\mathbf{H}(\mathbf{x})\|^{1/2}} K(\mathbf{H}(\mathbf{x})^{-1/2}(\mathbf{x} - \mathbf{x}_i))$$

where $\mathbf{H}(\mathbf{x})$ is the smoothing matrix for the estimation point $\mathbf{x}$. For each point at which the density is to be estimated, kernels of the same size and orientation are centered at each data point. The density estimate is computed by taking the average of the heights of the kernels at the estimation point. A popular choice for the bandwidth function in this case is to restrict the kernel to be spherically symmetric that further simplifies the approximation. Then, only one independent smoothing parameter remains $h_k(\mathbf{x})$ which is typically estimated as the distance from $\mathbf{x}$ to the kth nearest data point. Such an estimator suffers from several disadvantages - discontinuities, bias problems and integration to infinity.

An alternate strategy is to have the bandwidth matrix be a function of the sample points. Such estimator is called the *sample-point* estimator[24]:

$$\hat{f}_S(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}(\mathbf{x}_i)}(\mathbf{x} - \mathbf{x}_i))$$
$$= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\|\mathbf{H}(\mathbf{x}_i)\|^{1/2}} K(\mathbf{H}(\mathbf{x}_i)^{-1/2}(\mathbf{x} - \mathbf{x}_i))$$

The sample-point estimator still places a kernel at each data point. These kernels each have their own size and orientation regardless of where the density is to be estimated. This type of estimator was introduced by [4] who suggest using

$$\mathbf{H}(\mathbf{x}_i) = h(\mathbf{x}_i)\mathbf{I}$$

where $h(\mathbf{x}_i)$ is the distance from $\mathbf{x}_i$ to the k-th nearest data point. Asymptotically, this is equivalent to choosing $h(\mathbf{x}_i) \propto f(\mathbf{x}_i)^{-1/d}$ where d is the dimension of the data. A popular choice for the bandwidth function, suggested by [1], is to use $h(\mathbf{x}_i) \propto f(\mathbf{x}_i)^{-1/2}$ and, in practice, to use a pilot estimate of the density to calibrate the bandwidth function.

In this paper, we introduce a *hybrid* density estimator where the bandwidth is a function not only of the sample point but also of the estimation point $\mathbf{x}$. The particular property of the data that will be addressed is the existence of the uncertainty estimates of not only the sample points, but also the estimation point $\mathbf{x}$. Let $\{x_i\}_{i=1}^n$ be a set of measurements in d-dimensional space such that each $\mathbf{x}_i$ has associated with it a mean μ_i (in $\mathbb{R}^d$) and a $d \times d$ covariance matrix Σ_i. Also, let $\mathbf{x}$ (with mean $\mu_\mathbf{x}$ and covariance $\Sigma_\mathbf{x}$) be the current measurement whose probability is to be estimated. We define the multivariate *hybrid* density estimator as:

$$\hat{f}_H(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}(\mathbf{x},\mathbf{x}_i)}(\mu - \mu_i))$$
$$= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\|\mathbf{H}(\mathbf{x},\mathbf{x}_i)\|^{1/2}} K(\mathbf{H}(\mathbf{x},\mathbf{x}_i)^{-1/2}(\mu - \mu_i)) \quad (2)$$

where the bandwidth matrix $\mathbf{H}(\mathbf{x},\mathbf{x}_i)$ is a function of both the *estimation* measurement $\mathbf{x}$ and the *sample* measurement $\mathbf{x}_i$. Chen et. al. [5] have suggested using $\mathbf{H}_i = \chi^2_{\gamma,p}\Sigma_i$ for Epanechnikov kernels in the absence of error measurements. Expanding this idea, we propose the use of $\mathbf{H}(\mathbf{x},\mathbf{x}_i) = \Sigma_{\mathbf{x}_i} + \Sigma_\mathbf{x}$ as a possible bandwidth matrix for

the *Normal* kernel. Thus, the density estimator becomes [3] :

$$\hat{f}_H(\mathbf{x}) = \frac{1}{n(2\pi)^{d/2}} \sum_{i=1}^{n} \frac{1}{\|\Sigma_{\mathbf{x}_i} + \Sigma_{\mathbf{x}}\|^{1/2}} exp\left(-\frac{1}{2}(\mu - \mu_i)^T(\Sigma_{\mathbf{x}_i} + \Sigma_{\mathbf{x}})^{-1}(\mu - \mu_i)\right) \quad (3)$$

This particular choice for the bandwidth function has a simple but meaningful mathematical foundation. Suppose $\mathbf{x}_1$ and $\mathbf{x}_2$ are two normally distributed random variables with means $\{\mu_i\}$ and covariance matrices $\{\mathbf{\Sigma}_i\}$, i.e. $\mathbf{x}_i \sim N(\mu_i, \mathbf{\Sigma}_i), i = 1, 2$. It is well-known that if $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent, the distribution of $(\mathbf{x}_1 - \mathbf{x}_2)$ is $N(\mu_1 - \mu_2, \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)$. Thus, the probability that $\mathbf{x}_1 = \mathbf{x}_2$ or $\mathbf{x}_1 - \mathbf{x}_2 = \mathbf{O}$ is

$$p(\mathbf{x}_1 = \mathbf{x}_2) = \frac{1}{(2\pi)^{d/2}\|\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2\|^{1/2}} exp\left(-\frac{1}{2}(\mu_1 - \mu_2)^T(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-1}(\mu_1 - \mu_2)\right)$$

Thus, Equation 3 can be thought of as the average of the probabilities that the estimation measurement is equal to the sample measurement, calculated over all the sample measurements.

The choice for the bandwidth matrix can also be justified by the fact that the directions in which there is more uncertainty are given proportionately less weightage. Such uncertainty can be either in the estimation measurement or the sample measurements. Experimentally, the results obtained using these criteria were satisfactory when compared with the fixed bandwidth estimator or the balloon/sample-point estimators.

3 Classification

Once an appropriate mechanism for density approximation is built, the next step is to determine a classification mechanism for the observed data. Classification may be performed by thresholding on the probability of a new observation to belong to the background. However, two observations need to be taken into account:

- The threshold should be adaptive and determined based on the uncertainty or spread of the background distribution at a particular pixel (called *entropy* in information theory).
- Any available prior information about the foreground distribution should be utilized.

[3] Given the sample measurements, this estimator is a function of both the mean and the covariance of the estimation measurement. Thus, it is not a density function in the traditional sense which is a function only of the estimation point in d-dimensional space. However, if the covariance matrix $\Sigma_{\mathbf{x}}$ is kept fixed, it becomes a proper density function.

Figure 1. Adaptive thresholds for (a) The Ocean Sequence, (b) Traffic Sequence. Notice that the thresholds are higher in regions of low variability and low in regions of high variability.

More formally, guaranting a false-alarm rate of less than α_f requires that the threshold T should be set such that:

$$\int_{\hat{f}(\mathbf{x})<T} \hat{f}(\mathbf{x})d\mathbf{x} < \alpha_f \quad (4)$$

Furthermore, if $f_o(\mathbf{x})$ is the foreground distribution, guaranteeing a *miss* probability of α_m leads to the following condition on T:

$$\int_{\hat{f}_o(\mathbf{x})>T} \hat{f}_o(\mathbf{x})d\mathbf{x} > \alpha_m \quad (5)$$

Meeting both constraints simultaneously could be impossible, therefore a compromise is generally required. Furthermore, the foreground distribution is generally unknown weakening the use of the second constraint.

Determination of the threshold according to Equation (4) involves the inversion of complex integrals of clipped distributions. Such solution is feasible only for simple distributions like the Gaussian [10]. In the presence of more complex underlying densities, a statistical approximation is more suitable. We propose the use of sampling to get an estimate of the false alarm rate for a given threshold. Samples are drawn from the learnt background distribution (estimated via kernels in the present work) and the density at these sample points is classified using the current threshold as background or foreground. These classifications provide an estimate of the false alarm rate for the current threshold value. Such information can then be utilized for adjusting the threshold according to the desired false alarm rate. Since the "spread" of the distribution at a particular pixel is not expected to vary significantly over time, such threshold can be adjusted incrementally. Incremental adaptation of the threshold reduces the false alarms in the regions of high variation (e.g. waves, trees) while maintaining high detection rates in stationary areas.

4 Measurement of Features and their Uncertainties

Once the appropriate generic model for background subtraction is introduced, addressing the selection/estimation of the features is to be considered. As mentioned earlier, we utilize five features - two for optical flow and three for the intensity in the normalized color space. We have assumed that the uncertainties in their measurements are available. Here, we briefly describe methods that might be used for obtaining such measurements and their associated uncertainties.

4.1 Optical Flow

Several optical flow algorithms and their extensions [22, 18, 14, 6, 2] can be considered[4]. The most suitable for our approach[22] is briefly described next. It has the desired characteristic of being able to determine the error characteristics of the optical flow estimate. The method proposed by Simoncelli [22], is an extension to the method of Lucas and Kanade [18]. The basic idea of this algorithm is to apply the optical flow constraint equation [14]:

$$\nabla^T g.f + g_t = 0$$

where ∇g and g_t are the spatial image gradient and temporal derivative, respectively, of the image at a given spatial location and time, and f is the two-dimensional velocity vector. Such equation puts only one constraint on the two parameters (*aperture problem*). Thus, a smoothness constraint on the field of velocity vectors is a common selection to address this limitation. If we assume locally constant velocity and combine linear constraints over local spatial regions, a sum-of-squares error function can be defined:

$$E(f) = \sum_i w_i[\nabla^T g(x_i,t)f + g_t(x_i,t)]^2$$

Minimizing this error function with respect to f yields:

$$f = -M^{-1}b$$

where

$$M = \sum \nabla g \nabla^T g = \begin{bmatrix} \sum g_x^2 & \sum g_x g_y \\ \sum g_x g_y & \sum g_y^2 \end{bmatrix},$$

$$b = \begin{bmatrix} \sum g_x g_t \\ \sum g_y g_t \end{bmatrix} \tag{6}$$

and all the summations are over a patch around the point.

In [22], a model for recovering the uncertainties is introduced in the following way. Define $\hat{f}$ as the optical flow, f as the actual velocity field, and n_1 as the random variable describing the difference between the two. Then:

$$\hat{f} = f + n_1$$

Similarly, let $\hat{g}_t$ be the actual temporal derivative, and g_t the measured derivative. Then:

$$g_t = \hat{g}_t + n_2$$

where n_2 is a random variable characterizing the uncertainty in this measurement relative to the true derivative. The uncertainty in the spatial derivatives is assumed to be much smaller than the uncertainty in the temporal derivatives.

Under the assumption that n_1 and n_2 are governed by a normal distribution with covariance matrices $\mathbf{\Lambda}_1 = \lambda_1 \mathbf{I}$ and $\Lambda_2 = \lambda_2$ (it is scalar), and the flow vector $\mathbf{f}$ has a zero-mean Normal prior distribution with covariance $\mathbf{\Lambda}_p$, the covariance and mean of the optical flow vector may be estimated:

$$\begin{aligned} \Lambda_f &= \left[\sum_i \frac{w_i \mathbf{M}_i}{(\lambda_1 ||\nabla g(x_i)||^2 + \lambda_2)} + \Lambda_p^{-1} \right]^{-1} \\ \mu_f &= -\Lambda_f \sum_i \frac{w_i \mathbf{b}_i}{(\lambda_1 ||\nabla g(x_i)||^2 + \lambda_2)} \end{aligned} \tag{7}$$

where w_i is a weighting function over the patch, with the points in the patch indexed by i, and $\mathbf{M}_i$, and $\mathbf{b}_i$ are the same as matrices defined in Equation 6 but without the summation and evaluated at location x_i.

In order to handle significant displacements, a multiscale approach is considered that uses the flow estimates from a higher scale to initialize the flow for a lower level. Towards the propagation of variance across scales, a kalman filter is used with the normally used time variable replaced by scale. Further details of the approach may be obtained from the original paper by Simoncelli [22].

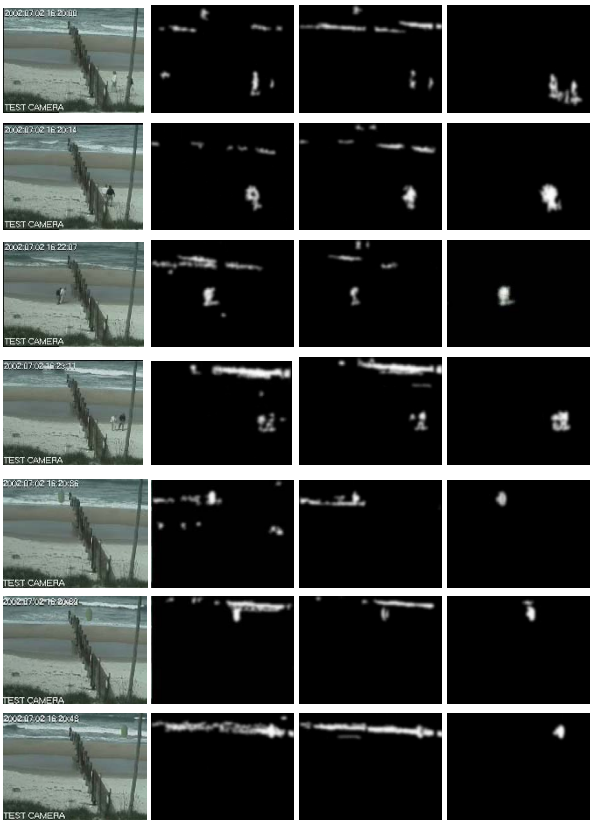
4.2 Normalized Color Representation

Suppose R,G and B are the RGB values observed at a pixel. Then, the normalized features are defined as:

$$r = R/S, \quad g = G/S, \quad I = S/3 \tag{8}$$

where $S = R + G + B$. The advantage of such transformation is that, under certain conditions, it is invariant to a change in illumination. However, such transformation introduces *heteroscedastic* (point-dependent) noise in the data that needs to be modeled correctly. Assuming that the sensor noise (in RGB space) is normally distributed with a diagonal covariance matrix having diagonal terms σ, it is not too difficult to show [11] that the uncertainties in the

[4] A survey and performance analysis of existing methods is presented in [3].



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

Figure 2. (a) Original Images. Detection result using (b) a mixture of Gaussians model (c) a non-parametric model, and (d) our approach. Simple spatial and temporal filtering was used for all algorithms.

normalized features is:

$$\Sigma_{r,g} = \frac{\sigma^2}{S^2} \begin{pmatrix} \left(1 - \frac{2R}{S} + \frac{3R^2}{S^2}\right) & \left(-\frac{R+G}{S} + \frac{3RG}{S^2}\right) \\ \left(-\frac{R+G}{S} + \frac{3RG}{S^2}\right) & \left(1 - \frac{2G}{S} + \frac{3G^2}{S^2}\right) \end{pmatrix}$$

4.3 Combining the Features

The covariance Σ_i for an observation $\mathbf{x}_i$ (in 5D space) may be estimated from the covariances of the components - the normalized color and optical flow. Assuming that the intensity and optical flow features are uncorrelated (which may not be true in general), an expression for the covariance matrix may be derived:

$$\Sigma_i = \begin{bmatrix} \Sigma_{\hat{r},\hat{g}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_i & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Lambda_f \end{bmatrix} \quad (9)$$

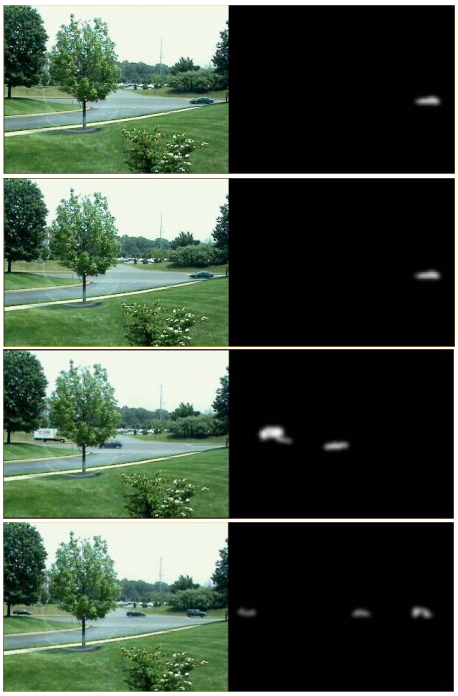


Figure 3. Detection for a traffic sequence consisting of waving trees.

where Λ_f is obtained as in Equation 7 and boldface $\mathbf{O}$'s represent appropriate zero matrices.

5 Results and Conclusion

5.1 Experimental Results

In order to validate the proposed technique, two different types of scenes were considered. The first is the challenging scene of the ocean front. Such scene involves wave motion, blowing grass, long-term changes due to tides, global illumination changes, shadows etc. An assessment of the performance of the existing methods [12, 8] is shown in Figure [2]. Even though these techniques were able to cope to some extent with the appearance change of the scene, their performance can be considered unsatisfactory for video based surveillance systems. The detection of events was either associated with a non-acceptable false alarm rate or the detection was compromised when focus was given to reducing the false alarm rate.

On the other hand, our algorithm was able to detect events of interest in the land and simulated events on the ocean front with extremely low false alarm rate as shown in Figure [5]. Large-scale experiments were conducted for this scene using several days of videos. Over this period, there were only 4 false alarms, occurring due to the reflection of

moving clouds in the water. On the other hand, the algorithm was able to detect simulated objects having almost no visual difference from the ocean if they were moving in a pattern that was different from the ocean [Figure 5 (i) - (l)]. At the same time, the algorithm had superior performance in the static parts of the scene because of the additional use of the optical flow component. This performance is reflected in the *ROC* curves for the various methods for this sequence [Figure 4(a)].

A typical traffic surveillance scenario was considered next where the challenge was due to the vigorous motion of the trees and bushes [Figure 3]. Again, our method was able to deal with the motion of the trees and outperformed existing techniques [Figure 4(b)]. Movie files containing some of these results have been made available on the conference website.

Since the information needs to be stored and evaluated for a sufficiently large temporal window, a limitation of the approach was its high computational and storage needs. Several optimizations can, however, be performed to reduce such requirements and our implementation is already running at about 7fps on a 160×120 3-band video on a Pentium IV 3GHz processor using about 400 MB of RAM for a temporal window size of 200 frames.

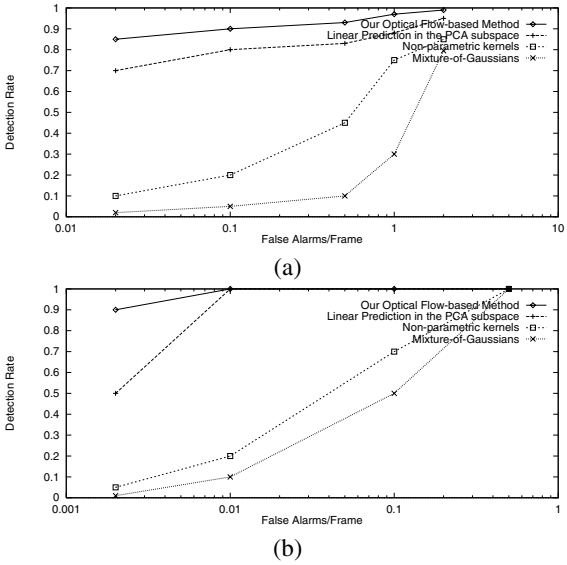


Figure 4. *Receiver-Operator Characteristic (ROC)* curves for (a) "ocean" sequence and (b) "traffic" sequence for (i) Mixture-of-Gaussians model [12], (ii) Non-parametric Kernels [8], (iii) Linear Prediction in PCA subspace [20], and (iv) Our method.

5.2 Discussion

In this paper we have proposed a technique for the modeling of dynamic scenes for the purpose of background-foreground differentiation and change detection [19]. The method relies on the utilization of optical flow as a feature for change detection. In order to properly utilize the uncertainties in the features, we proposed a novel kernel-based multivariate density estimation technique that adapts the bandwidth according the uncertainties in the test and sample measurements.

The algorithm had satisfactory performance in challenging settings. Detection performance was a function of the complexity of the observed scene. High variation in the observation space reflected to a mechanism with limited discrimination power. The method was able to adapt with global and local illumination changes, weather changes and changes of the natural scene.

We are investigating the evaluation of correlation that may exist between the different features in a multi-dimensional space. Also to be investigated is the use of other features like edges. Such incorporation will again require the proper evaluation of uncertainties and their correlation with other features. Last but not least, one can consider the modeling of scenes that exhibit more complex patterns of dynamic behavior. Such scenes may, for example, exhibit some dependencies between neighboring pixels. More sophisticated tools that take decisions at a higher level and are able to represent more sophisticated patterns of dynamic behavior is an interesting topic for further research.

Acknowledgements

We would like to thank Dorin Comaniciu for helpful discussions on the detection measure in a higher dimensional space, and Ramesh Visvanathan for pointing out the applicability of Simoncelli's method to the problem and for his support for conducting this work.

References

[1] I. Abramson. On bandwidth variation in kernel estimates- a square root law. *The Annals of Statistics*, 10:1217–1223, 1982.

[2] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *IJCV*, 2(3):283–310, January 1989.

[3] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *IJCV*, 12(1):43–77, February 1994.

[4] L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19:135–144, 1977.

[5] H. Chen and P. Meer. Robust computer vision through kernel density estimation. In *ECCV*, pages I: 236–250, Copenhagen, Denmark, May 2002.

[6] D. Comaniciu. Nonparametric information fusion for motion estimation. In *CVPR*, Madison, Wisconsin, June 2003.

[7] G. Doretto, A. Chiuso, Y.N. Wu, and S. Soatto. Dynamic textures. *IJCV*, 51(2):91–109, February 2003.

[8] A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background subtraction. In *ECCV*, pages II:751–767, Dublin, Ireland, May 2000.

[9] N. Friedman and S. Russell. Image segmentation in video sequences: A probabilistic approach. In *Thirteenth Conference on Uncertainty in Artificial Intelligence(UAI)*, August 1997.

[10] X. Gao, T.E. Boult, F. Coetzee, and V. Ramesh. Error analysis of background adaption. In *CVPR*, pages I: 503–510, Hilton Head Island, SC, June 2000.

[11] M. Greiffenhagen, V. Ramesh, D. Comaniciu, and H. Niemann. Statistical modeling and performance characterization of a real-time dual camera surveillance system. In *CVPR*, pages II:335–342, Hilton Head, SC, 2000.

[12] W.E.L. Grimson, C. Stauffer, R. Romano, and L. Lee. Using adaptive tracking to classify and monitor activities in a site. In *CVPR*, Santa Barbara, CA, June 1998.

[13] M. Harville. A framework for high-level feedback to adaptive, per-pixel, mixture-of-gaussian background models. In *ECCV*, page III: 543 ff., Copenhagen, Denmark, May 2002.

[14] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, pages 17:185–203, 1981.

[15] O. Javed, K. Shafique, and M. Shah. A hierarchical approach to robust background subtraction using color and gradient information. In *MVC*, pages 22–27, Florida, December 2002.

[16] Klaus-Peter Karmann and Achim von Brandt. *V Cappellini (ed.), Time Varying Image Processing and Moving Object Recognition*, volume 2, chapter Moving Object Recognition Using an Adaptive Background Memory. Elsevier, Amsterdam, The Netherlands, 1990.

[17] Dieter Koller, Joseph Weber, and Jitendra Malik. Robust multiple car tracking with occlusion reasoning. In *ECCV*, pages 189–196, Stockholm, Sweden, May 1994.

[18] B.D. Lucas and T. Kanade. An iterative technique of image registration and its application to stereo. In *Proc. 7th Int'l Joint Conf. on Artificial Intelligence*, 1981.

[19] A. Mittal, N. Paragios, V. Ramesh, and A. Monnet. A method for scene modeling and change detection. US Patent pending, 2003.

[20] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh. Background modeling and subtraction of dynamic scenes. In *ICCV*, pages 1305–1312, Nice, France, October 2003.

[21] N.M. Oliver, B. Rosario, and A.P. Pentland. A bayesian computer vision system for modeling human interactions. *PAMI*, 22(8):831–843, August 2000.

[22] E.P. Simoncelli. Bayesian multi-scale differential optical flow. In *Handbook of Computer Vision and Applications, Academic Press*, volume 2, pages 397–422, 1999.

[23] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *ICCV*, pages 255–261, Kerkyra, Greece, September 1999.

[24] M.P. Wand and M.C. Jones. *Kernel Smoothing*. Chapman and Hall, 1995.

[25] C.R. Wren, A. Azarbayejani, T.J. Darrell, and A.P. Pentland. Pfinder: Real-time tracking of the human body. *PAMI*, 19(7):780–785, July 1997.

[26] J. Zhong and S. Sclaroff. Segmenting foreground objects from a dynamic, textured background via a robust kalman filter. In *ICCV*, pages 44–50, Nice, France, October 2003.

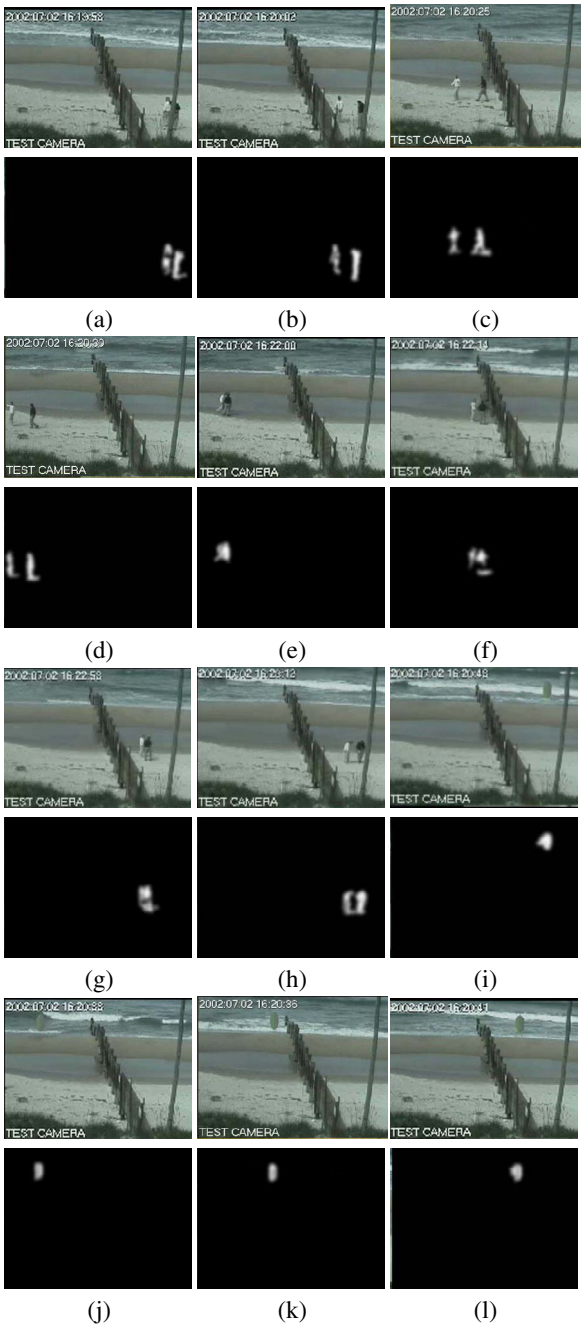


Figure 5. Additional results for the ocean sequence using the proposed algorithm. Note that, in Figures (i) - (l), an object with the same visual properties as the ocean was detected because of exhibiting different motion characteristics (horizontal).