Background Modeling and Subtraction of Dynamic Scenes

Antoine Monnet Anurag Mittal Nikos Paragios Visvanathan Ramesh
Real-Time Vision and Modeling
Siemens Corporate Research
755 College Road East, Princeton, NJ 08540, USA
e-mail: {anurag, nikos, rameshv }@scr.siemens.com

Abstract

Background modeling and subtraction is a core component in motion analysis. The central idea behind such module is to create a probabilistic representation of the static scene that is compared with the current input to perform subtraction. Such approach is efficient when the scene to be modeled refers to a static structure with limited perturbation.

In this paper, we address the problem of modeling dynamic scenes where the assumption of a static background is not valid. Waving trees, beaches, escalators, natural scenes with rain or snow are examples. Inspired by the work proposed in [4], we propose an on-line auto-regressive model to capture and predict the behavior of such scenes. Towards detection of events we introduce a new metric that is based on a state-driven comparison between the prediction and the actual frame. Promising results demonstrate the potentials of the proposed framework.

1 Introduction

The proliferation of cheap sensors and increased processing power has made the acquisition and processing of video information more feasible. Real-time video analysis tasks such as object detection and tracking can increasingly be performed efficiently on standard PC's for a variety of applications such as: Industrial automation, transportation, automotive, security & surveillance, and communications. The use of stationary cameras is rather common in several applications.

Background subtraction is a core component in such applications where the objective is to separate the foreground from the static parts of the scene. The information provided by such a module can be considered as a valuable low-level visual cue to perform high-level tasks of motion analysis, like motion estimation, tracking, etc. To this end, one has to obtain a representation of the background, update this representation over time and compare it with the actual input to determine areas of discrepancy.

Such methods have to be adaptive and able to deal with changes of the illumination conditions. Image averaging over a certain window of time is a computationally efficient approach to provide a fair description of the static scene in the absence of moving objects. A step further involves the use of continuous functions to better describe the illumination behavior of such a scene. Under the assumption of limited and smooth variation, in [19] a Kalman-filter driven approach was proposed to capture the background properties while in [26] the use of a single Gaussian distribution was considered.

The use of multiple hypotheses to describe the behavior of an evolving scene at the pixel level [10] was a breakthrough in the area of background modeling and subtraction. Such an approach is capable of dealing with significant variations, and was to be the basis for a large number of related techniques [14, 11, 12]. Parametric methods are a reasonable compromise between low complexity and fair approximation of the signal when it obeys the general assumptions imposed by the selected model. To deal with such limitation, in [5] a non-parametric approach was proposed. Their model was capable of describing the background density using temporal samples of the intensity at the pixel level.

The methods presented in [10, 5, 6, 24, 16] can effectively describe scenes that have a smooth behavior and limited variation. Consequently, they are able to cope with gradually evolving scenes. However, their performance deteriorates [Figure (5)] when the scene to be described is dynamic and exhibits non-stationary properties in time. Examples of such scenes are shown [Figure (3)] and include ocean waves, waving trees, rain, moving clouds, etc. Such events refer to a consistent pattern of change of the observation space in the spatio-temporal domain.

In this paper, we present a method for background modeling that is able to account for dynamic scenes. Using the ideas proposed in [4, 22], we treat the image as a time series and consider a predictive model to capture the most important variation based on a sub-space analysis of the signal. The components of this model are used in an auto-regressive form to predict the frame to be observed. Differ-

ences in the state space between the prediction and the observation quantify the amount of change and are considered to perform detection. Two different techniques are studied to maintain the model, one that update the states in an incremental manner and one that replaces the modes of variation using the latest observation map.

The reminder of the paper is organized as follows: in section 2, we briefly present the concept of the method while in section 3, we discuss the construction and maintenence of the background model. The detection mechanism is considered in section 4. Experimental results and discussion appear in section 5.

2 Scene Modeling

Let $\{\mathbf{I}(t)\}_{t=1\ldots\tau}$ be a given set of images [1]. The central idea behind our approach is to generate a prediction mechanism that can determine the actual frame using the k latest observed images. Such an objective can be defined in a more rigorous mathematical formulation as follows:

$$\mathbf{I}_{pred}(t) = f(\mathbf{I}(t-1), \mathbf{I}(t-2), \ldots, \mathbf{I}(t-k)) \quad (1)$$

where f, a k-th order function is to be determined. Quite often, information provided by the input images is rather complex and cannot be used in an efficient manner for prediction. Furthermore, solving the inference problem leads to a high-dimensional search space. In that case, complex techniques and significant amount of samples are required in order to recover a meaningful solution. One can address this limitation by the use of spatial filter operators. Complexity reduction of the search space and efficient data representation are the outcome of such procedure. Let $\{\phi_i\}_{i=1}^{n}$, be a filter bank and $\mathbf{s}_i(t) = \phi_i(\mathbf{I}(t))$, the output of convolution between the operator ϕ_i and the image $\mathbf{I}(t)$. The outcome of such a convolution process can be combined into a vector that represents the current state $\mathbf{s}(t)$ of the system.

$$\mathbf{s}^T(t) = [s_1(t), \ldots, s_n(t)]$$

Wavelet operators, Gabor filters, anisotropic non-linear filters can be considered. Within the proposed framework, linear filters are considered. Limited complexity and existence of efficient algorithms are the main motivation for such selection. Moreover, such filters are able to capture a significant amount of variations in real scenes.

2.1 Feature space

Based on the predictive model that was earlier introduced (Equation 1), a similar concept can be defined in the state space. Similar notation can be considered leading to

$$\mathbf{s}_{pred}(t) = f(\mathbf{s}(t-1), \mathbf{s}(t-2), \ldots, \mathbf{s}(t-k)) \quad (2)$$

[1] In order to achieve insensitivity to changes in illumination, the input may be transformed into a normalized space of $S = R+G+B, r = R/S$, and $g = G/S$, and the quantity S divided by an appropriate constant.

Several techniques can be considered to determine such prediction model. Principal component analysis [13, 3] refers to a linear transformation of variables that retains - for a given number n of operators - the largest amount of variation within the training data. In a prediction mechanism, such module can retain and recover in an incremental manner the core variations of the observed data.

The estimation of such operators will be addressed in the next section. In order to facilitate the introduction of the method at a concept level, one can consider them known $\{\phi_i = \mathbf{b}_i\}_{i=1}^{n}$, where $\mathbf{b}_i$ are the set of basis vectors. We consider these operators to produce the state vector $\mathbf{s}(t)$:

$$\begin{aligned}\mathbf{s}(t) &= [\phi_1(\tilde{\mathbf{I}}(t)), \phi_2(\tilde{\mathbf{I}}(t)), \ldots, \phi_n(\tilde{\mathbf{I}}(t))]^T \\ &= [\mathbf{b}_1^T \cdot \tilde{\mathbf{I}}(t), \mathbf{b}_2^T \cdot \tilde{\mathbf{I}}(t), \ldots, \mathbf{b}_n^T \cdot \tilde{\mathbf{I}}(t)]^T \\ &= \mathbf{B}^T \cdot \tilde{\mathbf{I}}(t)\end{aligned}$$

where $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_n]$ denotes the matrix of basis vectors, and $\tilde{\mathbf{I}}(t) = \mathbf{I}(t) - \bar{\mathbf{I}}$ denotes the mean subtracted input.

2.2 Prediction mechanism

The next step - given the selected feature space - refers to the modeling and estimation of the prediction function f. One can consider various forms (linear or non-linear) for such prediction mechanism. Non-linear mechanisms involve higher sophistication and can capture more complicated structures. However, the estimation of such functions is computationally expensive and suffers from instability.

Linear models are a good compromise between low complexity and a fairly good approximation of the observed structure. Auto-regressive models of a certain order k can be considered to approximate and predict the actual observation based on the latest k feature vectors. The predicted state will be a linear combination of these vectors:

$$\begin{aligned}\mathbf{s}_{pred}(t) &= f(\mathbf{s}(t-1), (\mathbf{s}(t-2), \ldots, (\mathbf{s}(t-k)) \\ &= \sum_{i=1}^{k} \mathbf{A}_i \mathbf{s}(t-i)\end{aligned}$$

where $\mathbf{A}$ is an $n \times n$ prediction matrix. The prediction in the image space can then be computed using the pseudo-inverse of $\mathbf{B^T}$:

$$\tilde{\mathbf{I}}_{pred}(t) = pseudoinv(\mathbf{B}^T) \cdot \mathbf{s}_{pred}(t)$$

where the pseudo-inverse is defined as

$$pseudoinv(\mathbf{B}^T) = (\mathbf{B}\mathbf{B}^T)^{-1} \cdot \mathbf{B} = \mathbf{B}$$

since $\mathbf{B}\mathbf{B}^T = \mathbf{I}$ (The basis vectors are orthogonal and have unit norm). Thus, the unknown variables of our scene model consist of the basis vectors and the auto-regressive matrix.

Visual appearance of indoors/outdoors scenes evolves over time. Global and local illumination changes, position

of the light sources, tidal changes, etc. are some examples of such dynamic behavior. One can account for such changes by continuously updating both the basis vectors and the predictive model according to the changes on the observed scene. Last, but not least the discriminability of the model should be preserved in order to perform accurate detection.

3 Model Construction

Estimation of the basis vectors from the observed data set can be performed through singular value decomposition. Then one can update such estimation as follows: (i) Considering an observation set that consists of the last m frames and recomputing the SVD based on the new data, (ii) Performing an incremental update of the basis vectors with exponential forgetting where every new frame is used to revise the estimate of these vectors. Similar procedures can be considered when recovering the parameters of the auto-regressive model.

3.1 Estimation of Basis Vectors

3.1.1 Batch PCA

Let $\{\mathbf{I}(t)\}_{t=1\ldots m}$ be a column vector representation of the previous m observations. We assume that the dimensionaly of this vector is d. One can estimate the mean vector $\bar{\mathbf{I}}$ and subtract it from the input to obtain zero mean vectors $\{\tilde{\mathbf{I}}(t)\}$ Given the set of training examples and the mean vector, one can define the $d \times d$ covariance matrix:

$$\Sigma_{\tilde{\mathbf{I}}} = E\{\tilde{\mathbf{I}}(t)\tilde{\mathbf{I}}^T(t)\}$$

It is well known that the principal orthogonal directions of maximum variation for $\mathbf{I}(t)$ are the eigenvectors of $\Sigma_{\tilde{\mathbf{I}}}$[13]. Therefore, one can assume that the use of such vectors is an appropriate selection for the filter bank.

One can approximate $\Sigma_{\tilde{\mathbf{I}}}$ with the sample covariance matrix that is given by $\tilde{\mathbf{I}}_M\tilde{\mathbf{I}}_M^T$, where $\tilde{\mathbf{I}}_M$ is the matrix formed by concatenating the set of images $\{\tilde{\mathbf{I}}(t)\}_{t=1\ldots m}$. Then, the eigenvectors of $\Sigma_{\tilde{\mathbf{I}}}$ can be computed through the singular value decomposition (SVD) of $\tilde{\mathbf{I}}_M$:

$$\tilde{\mathbf{I}}_M = \mathbf{U}D\mathbf{V}^T \tag{3}$$

The eigenvectors of the covariance matrix $\Sigma_{\tilde{\mathbf{I}}}$ are the columns of the matrix $\mathbf{U}$ (reffered to as the basis vectors henceforth) while the elements of the diagonal matrix D are the square root of the corresponding eigenvalues and refer to the variance of the data in the direction of the basis vectors. Such information can be used to determine the number of basis vectors (n) required to retain a certain percentage of the variance in the data.

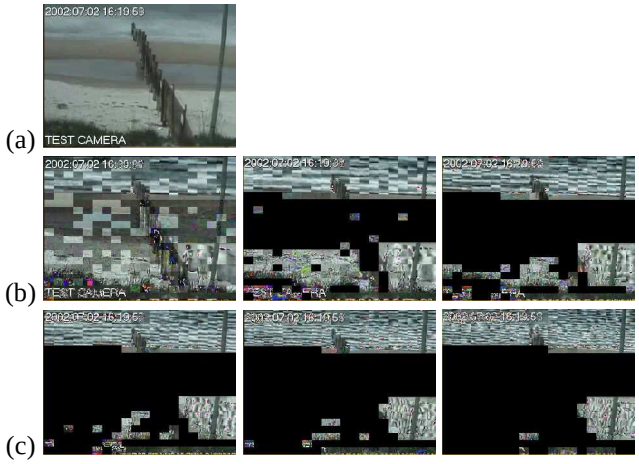


Figure 1. Basis Vectors: (a) mean, (b,c) mean + constant $\times$(6 principal) basis vectors. Insignificant basis vectors are represented with dark color.

Examples[2] of retained eigenvectors are shown in Figure (1). Information related with their magnitude and number are given in Figure (2).

3.1.2 Incremental PCA

The batch method is computationally inefficient and it might not be possible to execute it at each frame. Therefore, we consider a fast incremental method. The current estimate of the basis vectors is updated based on the new observation and the effect of the previous observations is exponentially reduced. Several methods for incremental PCA (IPCA) [25, 1] can be considered. We adapt the method developed by Weng et. al.[25] to suit to our application.

Amnesic Mean

Let $\mathbf{I}_1, \ldots \mathbf{I}_m$ be the previous m observations . The mean $\bar{\mathbf{I}}_m$ of the images can be computed incrementally:

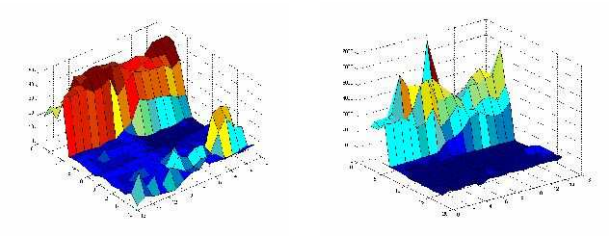
$$\bar{\mathbf{I}}_{m+1} = \left(\frac{m}{m+1}\right) \cdot \bar{\mathbf{I}}_m + \left(\frac{1}{m+1}\right) \cdot \mathbf{I}_{m+1}$$

This computation gives equal weight to all of the past observations. In order to reduce the effect of previous samples, one can compute the *amnesic* mean:

$$\bar{\mathbf{I}}_{m+1} = \left(\frac{m-l}{m+1}\right) \cdot \bar{\mathbf{I}}_m + \left(\frac{1+l}{m+1}\right) \cdot \mathbf{I}_{m+1}$$

where l is called the *amnesic* parameter that determines the rate of decay of the previous samples. If l is a fixed

[2]The image is divided into equal size blocks to reduce complexity.



(a) (b)

Figure 2. (a) Number of retained eigenvectors, (b) Magnitude of the largest eigenvalue.

multiple of m ($l = \lambda m$), one obtains exponential decay in the effect of past samples. Typical values of λ that we used were between 0.01 and 0.05.

Update of the Basis Vectors

Let $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ be the current set of estimates of the basis vectors. For reasons that become apparent later, these vectors are not normalized, although they are orthogonal [3]. Now, suppose we observe a new image $\mathbf{I}(t+1)$ and subtract the mean $\bar{\mathbf{I}}$ to obtain $\tilde{\mathbf{I}}(t+1)$. Then, we update the first basis vector $\mathbf{b}_1$ by essentially "pulling" it in the direction of $\tilde{\mathbf{I}}(t+1)$ by an amount equal to the projection of $\tilde{\mathbf{I}}(t+1)$ onto the unit vector along $\mathbf{b}_1$:

$$\mathbf{b}_1' = \left(\frac{m-l}{m+l}\right)\cdot\mathbf{b}_1 + \left(\frac{l+1}{t+1}\right)\left(\frac{\mathbf{b}_1 \cdot \tilde{\mathbf{I}}(\mathbf{t+1})}{\|\tilde{\mathbf{I}}(\mathbf{t+1})\|\|\mathbf{b_1}\|}\right)\cdot\tilde{\mathbf{I}}(t+1)$$

Here, $\frac{\mathbf{b}_1\cdot\tilde{\mathbf{I}}(\mathbf{t+1})}{\|\mathbf{b_1}\|}$ is the projection of $\tilde{\mathbf{I}}(t+1)$ in the direction of $\mathbf{b}_1$, and $\frac{\tilde{\mathbf{I}}(t+1)}{\|\tilde{\mathbf{I}}(t+1)\|}$ is the unit vector in the direction of $\tilde{\mathbf{I}}(t+1)$.

Next, we compute the residue $\mathbf{R}_1$ of $\tilde{\mathbf{I}}(t+1)$ on $\mathbf{b}_1$:

$$\begin{aligned}\mathbf{R}_1 &= \tilde{\mathbf{I}}(t+1) - Proj_{\mathbf{b}_1}(\tilde{\mathbf{I}}(t+1)) \\ &= \tilde{\mathbf{I}}(t+1) - \left(\frac{\mathbf{b}_1 \cdot \tilde{\mathbf{I}}(t+1)}{\|\mathbf{b}_1\|^2}\right)\cdot\mathbf{b}_1\end{aligned}$$

This residue is perpendicular to $\mathbf{b}_1$ and is used to "pull" $\mathbf{b}_2$ in the direction of $\mathbf{R}_1$ by an amount equal to the projection of $\mathbf{R}_1$ onto the unit vector along $\mathbf{b}_2$:

$$\mathbf{b}_2' = \left(\frac{m-l}{m+l}\right)\cdot\mathbf{b}_2 + \left(\frac{l+1}{t+1}\right)\left(\frac{\mathbf{b}_2 \cdot \mathbf{R}_1}{\|\mathbf{R}_1\|\|\mathbf{b}_2\|}\right)\cdot\mathbf{R}_1$$

The residue $\mathbf{R}_2$ is calculated similarly:

$$\begin{aligned}\mathbf{R}_2 &= \mathbf{R}_1 - Proj_{\mathbf{b}_2}(\mathbf{R}_1) \\ &= \mathbf{R}_1 - \left(\frac{\mathbf{b}_2 \cdot \mathbf{R}_1}{\|\mathbf{b}_2\|^2}\right)\cdot\mathbf{b}_2\end{aligned}$$

[3] However, they can be normalized for use in the background model.

This residue is perpendicular to the *span* of $< \mathbf{b}_1\mathbf{b}_2 >$. This procedure is repeated for each subsequent basis vector such that the basis vector $\mathbf{b}_j$ is pulled towards $\tilde{\mathbf{I}}(t+1)$ in a direction perpendicular to the span of $< \mathbf{b}_1 \ldots \mathbf{b}_{j-1} >$.

Zhang and Weng [25] have proved that, with the above algorithm, $\mathbf{b}_i \rightarrow \pm\lambda_i\mathbf{e}_i$ as $n \rightarrow \infty$. Here, λ_i is the i-th largest eigenvalue of the covariance matrix $\Sigma_{\tilde{\mathbf{I}}}$, and $\mathbf{e}_i$ is the corresponding eigenvector. Note that the obtained vector has a scale of λ_i and is not a unit vector. Therefore, in our application we store these unnormalized vectors. The magnitude λ_i yields the eigenvalue and normalization yields the eigenvector at any iteration.

3.2 Estimation of the predictive model

As stated earlier, we will use a linear auto-regressive model to model the transformation of states:

$$\mathbf{s}_{pred}(t) = \sum_{i=1}^{k} \mathbf{A}_i\mathbf{s}(t-i)$$

for a k-th order auto-regressive model. The parameters of the model, contained in $\mathbf{A}_i$'s, can be recovered if a data set of previous state transformations is available.

We illustrate the approach to be followed in computing the coefficients of the matrices $\mathbf{A}_i$ by considering the case of $k = 1$. Let us form two matrices from the state vectors $\{\mathbf{s}(t)\}$:

$$\mathbf{S2} = [\mathbf{s}(2), \mathbf{s}(3), \ldots, \mathbf{s}(t)]$$

$$\mathbf{S1} = [\mathbf{s}(1), \mathbf{s}(2), \ldots, \mathbf{s}(t-1)]$$

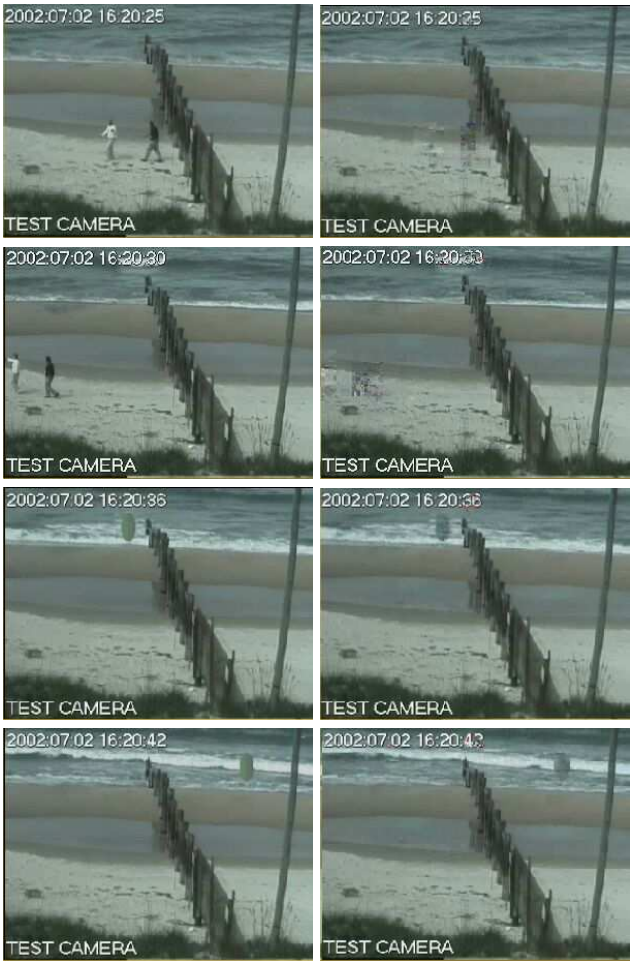
Now, the application of $\mathbf{s}_{pred}(t) = \sum_{i=1}^{k} \mathbf{A}_i\mathbf{s}(t-i) = \mathbf{A}_1\mathbf{s}(t-1)$ on the state transformations yields an over-constrained set of linear equations. This set of equations can then be solved for the best $\mathbf{A}$ in the sense of least squares error $argmin_{\mathbf{A}}\|\mathbf{S2} - \mathbf{A}\cdot\mathbf{S1}\|$ by the method of *normal equations* [4, 9]:

$$\mathbf{A} = \mathbf{S2}\cdot\mathbf{S1}^T\cdot(\mathbf{S1}\cdot\mathbf{S1}^T)^{-1}$$

A closed form solution for the optimal parameters of the auto-regressive model in a least squares sense is possible for any k. This is achieved by solving an over-constrained set of linear equations that are formed by taking each row of the equation $\mathbf{s}(t) = \sum_i \mathbf{A}_i\mathbf{s}(t-i)$, where the the coefficients of the $n \times n$ square matrices $\mathbf{A}_i$ are the unknowns. These can be solved using the method of *normal equations*.

4 Detection

A simple mechanism to perform detection is by comparing the prediction with the actual observation. Under the assumption that the auto-regressive model was built using background samples, such technique will provide poor



(a) (b)

Figure 3. (a) Input signal, (b) Prediction.

prediction for objects while being able to capture the background. Two types of changes in the signal may be considered for detection: (1) "structural" change in the appearance of pixel intensities in a given region, and (2) change in the motion characteristics of the signal. Measures are developed in order to detect each of these types of changes.

4.1 Structural Change

In order to develop the approach for estimating structural change in the signal, we begin by reviewing some concepts in Principal Component Analysis and its relationship to density estimation in a multi-dimensional space. The principal component analysis decomposes the vector space $\mathbb{R}^d$ into two mutually exclusive and complementary subspaces: the principal subspace (or feature space) $F = \{b_i\}_{i=1}^n$ containing the principal components and its orthogonal complement $\bar{F} = \{b_i\}_{i=n+1}^d$. Then, using $\mathbf{s} = \mathbf{B}^T \cdot \tilde{\mathbf{I}}$, the residual reconstruction error for an input vector $\tilde{\mathbf{I}}(t)$ can be defined as [7, 15]:

$$\epsilon^2(\tilde{\mathbf{I}}) = \sum_{i=n+1}^{d} s_i^2 = \|\tilde{\mathbf{I}}\|^2 - \sum_{i=1}^{n} s_i^2$$

This can be easily computed from the first n principal components and the L_2-norm of the mean-normalized image $\tilde{\mathbf{I}}$. Then, the L_2 norm of any element $\mathbf{x} \in \mathbb{R}^d$ can be decomposed in terms of its projection in these two subspaces. The component in the orthogonal subspace $\bar{F}$, referred to as the "distance-from-feature-space" (DFFS) [15], is a simple Euclidean distance equivalent to $\epsilon^2(\mathbf{x})$.

Let us assume a Gaussian model for the density in high-dimensional space. More complicated models for the density, like mixture-of-Gaussians, or non-parametric approaches can also be considered and easily integrated by explicitly building a background model on the state space. If we assume that the mean $\bar{\mathbf{I}}$ and covariance Σ of the distribution have been estimated robustly, the likelihood of an input $\mathbf{I}$ to belong to the background class Ω is given by:

$$p(\mathbf{I}|\Omega) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} exp\left(-\frac{1}{2}(\mathbf{I}-\bar{\mathbf{I}})^T\Sigma^{-1}(\mathbf{I}-\bar{\mathbf{I}})\right)$$

The sufficient statistic for characterizing this likelihood is the *Mahalanobis distance*:

$$d(\mathbf{I}) = \tilde{I}^T\Sigma^{-1}\tilde{I}$$

where $\tilde{\mathbf{I}} = \mathbf{I} - \bar{\mathbf{I}}$. Utilizing the eigenvalue decomposition of Σ: $\Sigma = \mathbf{B}\Lambda\mathbf{B}^T$ where $\mathbf{B}$ is the eigenvector matrix of Σ and is the same as the matrix of basis vectors $\mathbf{B}$ used earlier and Λ is the diagonal matrix of eigenvalues (= D^2 in Equation 3), the Mahalanobis distance can be written as:

$$d(\mathbf{I}) = \tilde{\mathbf{I}}^T\Sigma^{-1}\tilde{\mathbf{I}} = \tilde{\mathbf{I}}^T[\mathbf{B}\Lambda^{-1}\mathbf{B}^T]\tilde{\mathbf{I}} = \mathbf{s}^T\Lambda^{-1}\mathbf{s}$$

since $\mathbf{B}^T\tilde{\mathbf{I}} = \mathbf{s}$. Due to the diagonal form of Λ, we can rewrite this equation as:

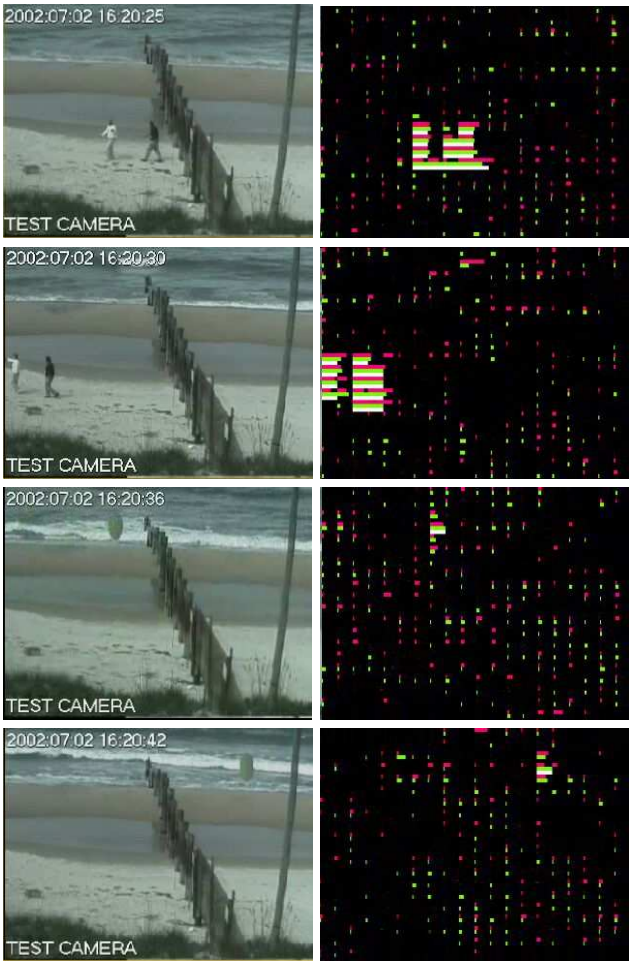
$$d(\mathbf{I}) = \sum_{i=1}^{d} \frac{s_i^2}{\lambda_i}$$

where λ_i is the i-th eigenvalue. If we seek to estimate $\tilde{d}(\mathbf{I})$ using only the n principal projections, one can formulate an optimum estimator for $\tilde{d}(\mathbf{I})$ as follows:

$$\begin{aligned}\tilde{d}(\mathbf{I}) &= \sum_{i=1}^{n} \frac{s_i^2}{\lambda_i} + \frac{1}{\rho}\left[\sum_{i=n+1}^{d} s_i^2\right] \\ &= \sum_{i=1}^{n} \frac{s_i^2}{\lambda_i} + \frac{1}{\rho}\epsilon^2(\tilde{\mathbf{I}})\end{aligned} \tag{4}$$

In [15], it was shown that an optimal ρ in terms of a suitable error measure based on the Kullback-Leibler divergence is:

$$\rho^* = \frac{1}{d-n}\sum_{i=n+1}^{d} \lambda_i$$



(a) (b)

Figure 4. (a) Input frames, (b) Detection Components. In each block, green represents r_1, pink shows r_2 and white represents detection by combining r_1 and r_2.

We propose the use of $\tilde{d}(\mathbf{I})$ as the first detection measure r_1. It is an optimum measure for estimating the distance from the Gaussian density represented by the principal component analysis such that the covariances of the data are properly taken into account while estimating the difference. High values of such distance measure have the following interpretation: the original vector is not close to the training data, and may correspond to a new object in the scene. In other words, this is a measure of change of the scene structure. Such case can occur either because of changes in the appearance of the scene (color), or because of structural changes. Therefore, such technique can better detect objects than the standard background subtraction techniques that consider each pixel individually without considering the relationships between them.

4.2 Change in Motion Characteristics

The measure r_1 can deal efficiently with changes of appearance in the structural sense but would fail to capture changes in the temporal domain. This can occur when information appears in a different temporal order than the one for the background. To this end, one can consider the *SSD* (sum of squared differences) error between the input and predicted image, which can be expressed as the square of the L_2 norm of the difference between the vectorized input and predicted images: $\|\mathbf{I} - \mathbf{I}_{pred}\|_2^2$. Since any vector $\mathbf{I}$ may be written in terms of its components along the basis vectors, $\mathbf{I} = \sum_{i=1}^{d} s_i \mathbf{B}_i$, we may write:

$$\mathbf{I} - \mathbf{I}^{pred} = \sum_{i=1}^{d} s_i \mathbf{B}_i - \sum_{i=1}^{d} s_i^{pred} \mathbf{B}_i = \sum_{i=1}^{d} (s_i - s_i^{pred}) \mathbf{B}_i$$

Therefore, the norm of this vector may be computed thus:

$$\|\mathbf{I} - \mathbf{I}_{pred}\|_2^2 = \|\sum_{i=1}^{d} (s_i - s_i^{pred}) \mathbf{B}_i\|_2^2 = \sum_{i=1}^{d} (s_i - s_i^{pred})^2$$
$$= \sum_{i=1}^{n} (s_i - s_i^{pred})^2 + \sum_{i=n+1}^{d} (s_i)^2$$

since the prediction is made from only the first n components, and therefore $s_i^{pred} = 0, i = n+1 \ldots d$. Recalling the definition of $\epsilon^2(\tilde{\mathbf{I}})$, we obtain:

$$\|\mathbf{I} - \mathbf{I}_{pred}\|_2^2 = \sum_{i=1}^{n} (s_i - s_i^{pred})^2 + \epsilon^2(\tilde{\mathbf{I}})$$

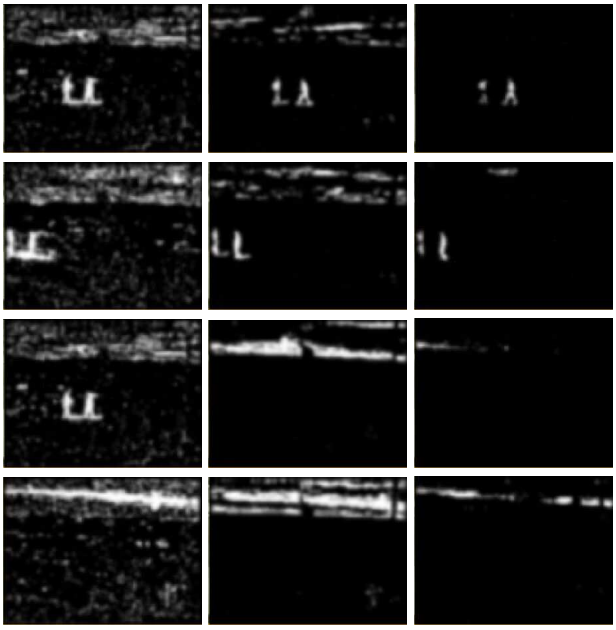
Since the effect of the second term has already been captured in r_1, we define

$$r_2(t) = \sum_{i=1}^{n} (s_i - s_i^{pred})^2 = \|\mathbf{s} - \mathbf{s}^{pred}\|_2^2$$

where the state vectors are considered only upto the n principal components. Such measure captures the change in the motion characteristics. Objects following motion trajectories different than the ones being considered (autoregressive model) will reflect to important values for r_2. Such metric is an additional cue for detection based on structural motion that has not been considered in traditional background adaptation methods [10, 5].

4.3 Implementation Details

Real-time processing is a standard requirement of video surveillance. In particular, when dealing with techniques that are aimed at background adaptation, such requirement is strictly enforced. Changes of the background structure should be captured from the model to preserve satisfactory detection rate.



(a) (b) (c)

Figure 5. Detection result for the images in Figure (4) using (a) a mixture of Gaussians model[10], (b) a non-parametric model[5], and (c) the non-parametric model with low detection threshold.

Computing the basis components for large vectors is a time consuming operation. Optimal algorithms for singular value decomposition of an $m \times n$ matrix take $O(m^2 n + n^3))$ time[9]. A simple way to deal with such complexity is by considering the process at a block level. To this end, we divide the image into blocks and run the algorithm independently on each block. For each of these blocks, the number of components retained is determined dynamically by the singular values (which refer to the standard deviation in the direction of basis vectors). Also determined by the singular values is the number of past images over which the SVD is computed (for the non-incremental method). This is because higher variation in a region suggests that more images would be required to model it. Furthermore, we compute the PCA only over those frames that are not well modeled by the current basis vectors. This enables us to capture the variation of the data over a much longer time window with the same computational cost.

Such mechanism leads to a quasi real-time ($\sim$5fps) implementation for a 340×240 3-band video stream on a 2.2 GHz Pentium 4 processor machine.

4.4 Experimental Results

In order to validate the proposed technique, the challenging scene of the ocean front was considered. Such scene involves wave motion, blowing grass, long-term changes due to tides, global illumination changes, shadows etc. An assessment on the performance of the existing methods [10, 5] is shown in Figure (5). 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. Our algorithm was able to detect events of interest in the land and simulated events on the ocean front as shown in Figure 4.

The essence of the approach is depicted in Figures (3) and (4). Observation as well as prediction are presented for comparison. Visually, one can conclude that the prediction is rather close to the actual observation for the background component. On the other hand, prediction quality deteriorates when a non-background structure appear in the scene. A more elaborate technique to validate prediction is through the detection process as shown in Figure (4).

Based on the experiments, one can claim that our approach was able to capture the dynamic structure of the ocean front as well as the blowing grass. At the same time, dealing with the static parts of the scene is trivial with the proposed framework.

A second scene we considered was a regular traffic surveillance scene that has several waving trees [Figure (6)]. The algorithm was again able to perform better than traditional methods and acheived reduced false alarm rate in the tree region without any manual adjustment of parameters.

5 Discussion

In this paper we have proposed a prediction-based on-line method for the modeling of dynamic scenes. The core contribution of our approach is the integration of a powerful set of filter operators within a linear prediction model towards the detection of events using measures that are adaptive to the complexity of the scene. Furthermore, we have proposed an on-line adaptation technique to maintain the selection of the best filter operators and the background model.

The approach has been tested and validated using a challenging setting, the detection of events on the coast line and the ocean front [Figure (3)]. Large scale experiments that involved real events as well as simulated ones were conducted. The proposed technique was able to detect such events with a minimal false alarm rate. 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, changes of the natural scene, etc. Our method could meet and overcome in some cases the performance of existing techniques [Figure (5)] for station-

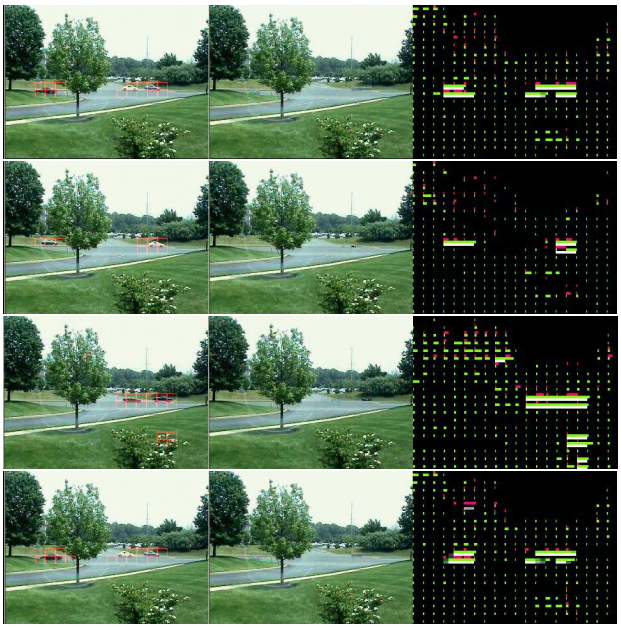


Figure 6. Results for a sequence of a road with waving trees. *Left*: Input signal, *Middle:* Predicted signal, *Right:* Block-wise response of the detection measures. As before, green represents r_1, pink represents r_2 and white represents detection in a block.

ary scenes, while being able to deal with more complex and evolving natural scenes.

On-line adaptation of the prediction mechanism (auto-regressive model) is an on-going work. Such component will further reduce the complexity and make the method more efficient. More sophisticated tools that take decisions at a higher level and consider neighborhood dependencies is something to be explored. Exploration of non-linear operators that can better capture the variation of the data leading to a strong discriminabilty for the model is a direction that has to be considered. Last, but not least more elaborated prediction mechanisms can further improve the performance of the algorithm.

Acknowledgements

This work was inspired by the work of Soatto et. al. [22, 4]. Moreover, they provided us with an implementation of their algorithm for which we are very grateful. We are also grateful to Silviu Minut for making available to us his implementation of Incremental PCA.

References

[1] M. Brand. Incremental singular value decomposition of uncertain data with missing values. In *ECCV*, page I: 707 ff., Copenhagen, Denmark, May 2002.

[2] R.T. Collins, A.J. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multi-sensor surveillance. *Proceedings of the IEEE*, 89(10):1456–1477, October 2001.

[3] F. de la Torre and M.J. Black. Robust principal component analysis for computer vision. In *ICCV*, pages I: 362–369, Vancouver, Canada, July 2001.

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

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

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

[7] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, Boston, second edition, 1990.

[8] 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.

[9] G. H. Golub and C.F. Van Loan. *Matrix Computations*. John Hopkins University Press, 1996.

[10] 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.

[11] 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.

[12] O. Javed, K. Shafi que, and M. Shah. A hierarchical approach to robust background subtraction using color and gradient information. In *IEEE Workshop on Motion and Video Computing*, pages 22–27, Orlando, Florida, December 2002.

[13] I.T. Jolliffe. *Principal Component Analysis*. Springler-Verlag, 1986.

[14] A. Mittal and D.P. Huttenlocher. Scene modeling for wide area surveillance and image synthesis. In *CVPR*, pages II: 160–167, Hilton Head, SC, 2000.

[15] B. Moghaddam and A.P. Pentland. Probabilistic visual learning for object representation. *PAMI*, 19(7):696–710, July 1997.

[16] N. Ohta. A statistical approach to background subtraction for surveillance systems. In *ICCV*, pages II: 481–486, Vancouver, Canada, June 2001.

[17] N. Paragios and V. Ramesh. A mrf-based approach for real-time subway monitoring. In *CVPR*, pages I:1034–1040, Hawaii, December 2001.

[18] Nikos Paragios and George Tziritas. Adaptive detection and localization of moving objects in image sequences. *Signal Processing: Image Communication*, 4:277–296, September 99.

[19] C. Ridder, O. Munkelt, and H. Kirchner. Adaptive background estimation and foreground detection using kalman fi ltering. In *Proc. International Conference on recent Advances in Mechatronics, 193–199*, 1995.

[20] A. Schodl, R. Szeliski, D. Salesin, and I. Essa. Video textures. In *Proceedings of ACM SIGGRAPH Conference*, New Orleans, LA, 2000.

[21] Makito Seki, Toshikazu Wada, H. Fujiwara, and K. Sumi. Background subtraction based on cooccurrence of image variations. In *CVPR*, Madison, WI, 2003.

[22] S. Soatto, G. Doretto, and Y.N. Wu. Dynamic textures. In *ICCV*, pages II: 439–446, Vancouver, Canada, July 2001.

[23] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, and J.M. Buhmann. Topology free hidden markov models: Application to background modeling. In *ICCV*, pages I: 294–301, Vancouver, Canada, 2001.

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

[25] J. Weng, Y. Zhang, and W.S. Hwang. Candid covariance-free incremental principal component analysis. *PAMI*, 25(8), 2003.

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