Bayesian Modelling of Dynamic Scenes for Object Detection

Yaser Sheikh and Mubarak Shah

Abstract

Accurate detection of moving objects is an important precursor to stable tracking or recognition. In this paper, we present an object detection scheme that has three innovations over existing approaches. Firstly, the model of the intensities of image pixels as independent random variables is challenged and it is asserted that useful correlation exists in intensities of spatially proximal pixels. This correlation is exploited to sustain high levels of detection accuracy in the presence of dynamic backgrounds. By using a non-parametric density estimation method over a joint domain-range representation of image pixels, multi-modal spatial uncertainties and complex dependencies between the domain (location) and range (color) are directly modeled. We propose a model of the background as a *single* probability density. Secondly, temporal persistence is proposed as a detection criterion. Unlike previous approaches to object detection which detect objects by building adaptive models of the background, the *foreground* is modeled to augment the detection of objects (without explicit tracking), since objects detected in the preceding frame contain substantial evidence for detection in the current frame. Finally, the background and foreground models are used competitively in a MAP-MRF decision framework, stressing spatial context as a condition of detecting interesting objects and the posterior function is maximized efficiently by finding the minimum cut of a capacitated graph. Experimental validation of the proposed method is performed and presented on a diverse set of dynamic scenes.

Keywords

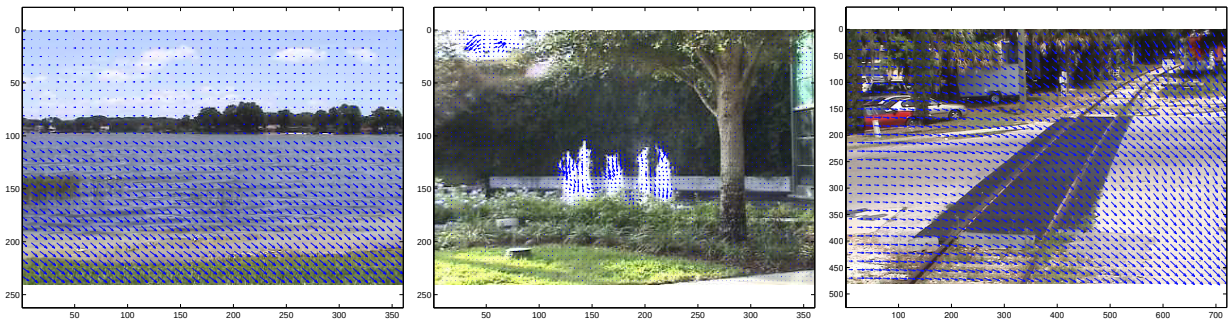
Object Detection, Kernel Density Estimation, Joint Domain Range, MAP-MRF Estimation.

I. Introduction

Automated surveillance systems typically use stationary sensors to monitor an environment of interest. The assumption that the sensor remains stationary between the

incidence of each video frame allows the use of statistical background modeling techniques for the detection of moving objects such as [39], [33] and [7]. Since 'interesting' objects in a scene are usually defined to be moving ones, such object detection provides a reliable foundation for other surveillance tasks like tracking ([14], [16], [5]) and is often also an important prerequisite for action or object recognition. However, the assumption of a stationary sensor does not necessarily imply a stationary *background*. Examples of 'nonstationary' background motion abound in the real world, including periodic motions, such as a ceiling fans, pendulums or escalators, and dynamic textures, such as fountains, swaying trees or ocean ripples (shown in Figure 1). Furthermore, the assumption that the sensor remains stationary is often *nominally* violated by common phenomena such as wind or ground vibrations and to a larger degree by (stationary) hand-held cameras. If natural scenes are to be modeled it is essential that object detection algorithms operate reliably in such circumstances. Background modeling techniques have also been used for foreground detection in pan-tilt-zoom cameras, [37]. Since the focal point does not change when a camera pans or tilts, planar-projective motion compensation can be performed to create a background mosaic model. Often, however, due to independently moving objects motion compensation may not be exact, and background modeling approaches that do not take such nominal misalignment into account usually perform poorly. Thus, a principal proposition in this work is that modeling spatial uncertainties is important for real world deployment, and we provide an intuitive and novel representation of the scene background that consistently yields high detection accuracy.

In addition, we propose a new constraint for object detection and demonstrate significant improvements in detection. The central criterion that is traditionally exploited for detecting moving objects is *background difference*, some examples being [17], [39], [26]



(a) (b) (c)

Fig. 1. Various sources of dynamic behavior. The flow vectors represent the motion in the scene. (a) The lake-side water ripples and shimmers (b) The fountain, like the lake-side water, is a temporal texture and does not have exactly repeating motion (c) a strong breeze can cause nominal motion (camera jitter) of upto 25 pixels between consecutive frames.

and [33]. When an object enters the field of view it partially occludes the background and can be detected through background differencing approaches if its appearance differs from the portion of the background it occludes. Sometimes, however, during the course of an object's journey across the field of view, some colors may be similar to those of the background, and in such cases detection using background differencing approaches fail. To address this limitation and to improve detection in general, a new criterion called *temporal persistence* is proposed here and exploited in conjunction with background difference for accurate detection. True foreground objects, as opposed to spurious noise, tend to maintain consistent colors and remain in the same spatial area (i.e. frame to frame color transformation and motion are small). Thus, foreground information from the frame incident at time t contains substantial evidence for the detection of foreground objects at time $t+1$. In this paper, this fact is exploited by maintaining both background and foreground models to be used competitively for object detection in stationary cameras, without explicit tracking.

Finally, once pixel-wise probabilities are obtained for belonging to the background, decisions are usually made by direct thresholding. Instead, we assert that *spatial context* is an important constraint when making decisions about a pixel label, i.e. a pixel's label is not independent of the pixel's neighborhood labels (this can be justified on Bayesian grounds using Markov Random Fields, [11], [23]). We introduce a MAP-MRF framework, that competitively uses both the background and the foreground models to make decisions based on spatial context. We demonstrate that the *maximum a posteriori* solution can be efficiently computed by finding the minimum cut of a capacitated graph, to make an optimal inference based on neighborhood information at each pixel.

The rest of the paper is organized as follows. Section I-A reviews related work in the field and discusses the proposed approach in the context of previous work. A description of the proposed approach is presented in Section I-B. In Section II, a discussion on modeling spatial uncertainty (Section II-A) and on utilizing the foreground model for object detection (Section II-B) and a description of the overall MAP-MRF framework is included (Section II-C). In Section II-C, we provide an algorithmic description of the proposed approach as well. Qualitative and quantitative experimental results are shown in Section III, followed by conclusions in Section IV.

A. Previous Work

Since the late 70s, differencing of adjacent frames in a video sequence has been used for object detection in stationary cameras, [17]. However, it was realized that straightforward background subtraction was unsuited to surveillance of real-world situations and statistical techniques were introduced to model the uncertainties of background pixel colors. In the context of this work, these background modeling methods can be classified into two categories: (1) Methods that employ *local* (pixel-wise) models of intensity and (2) Methods

that have *regional* models of intensity.

Most background modeling approaches tend to fall into the first category of pixel-wise models. Early approaches operated on the premise that the color of a pixel over time in a static scene could be modeled by a single Gaussian distribution, $N(\mu, \Sigma)$. In their seminal work, Wren *et al* [39] modeled the color of each pixel, $I(x, y)$, with a single 3 dimensional Gaussian, $I(x, y) \sim N(\mu(x, y), \Sigma(x, y))$. The mean $\mu(x, y)$ and the covariance $\Sigma(x, y)$, were learned from color observations in consecutive frames. Once the pixel-wise background model was derived, the likelihood of each incident pixel color could be computed and labelled as belonging to the background or not. Similar approaches that used Kalman Filtering for updating were proposed in [20] and [21]. A robust detection algorithm was also proposed in [14]. While these methods were among the first to principally model the uncertainty of each pixel color, it was quickly found that the single Gaussian *pdf* was ill-suited to most outdoor situations, since repetitive object motion, shadows or reflectance often caused multiple pixel colors to belong to the background at each pixel. To address some of these issues, Friedman and Russell, and independently Stauffer and Grimson, [9], [33] proposed modeling each pixel intensity as a *mixture* of Gaussians, instead, to account for the multi-modality of the 'underlying' likelihood function of the background color. An incident pixel was compared to every Gaussian density in the pixel's model and if a match (defined by threshold) was found, the mean and variance of the matched Gaussian density was updated, or otherwise a new Gaussian density with the mean equal to the current pixel color and some initial variance was introduced into the mixture. Thus, each pixel was classified depending on whether the matched distribution represented the background process. While the use of Gaussian mixture models was tested extensively, it did not explicitly model the *spatial dependencies* of neighboring pixel colors that may be

caused by a variety of real nominal motion. Since most of these phenomenon are 'periodic', the presence of multiple models describing each pixel mitigates this effect somewhat by allowing a mode for each periodically observed pixel intensity, however performance notably deteriorates since dynamic textures usually do not repeat exactly (see experiments in Section III). Another limitation of this approach is the need to specify the number of Gaussians (models), for the E-M algorithm or the K-means approximation. Still, the mixture of Gaussian approach has been widely adopted, becoming something of a standard in background subtraction, as well as a basis for other approaches ([18],[15]).

Methods that address the uncertainty of spatial location using local models have also been proposed. In [7], El Gammal *et al* proposed nonparametric estimation methods for per-pixel background modeling. Kernel density estimation (KDE) was used to establish membership, and since KDE is a data-driven process, multiple modes in the intensity of the background were also handled. They addressed the issue of nominally moving cameras with a local search for the best match for each incident pixel in neighboring models. Ren *et al* too explicitly addressed the issue of background subtraction in a nonstationary scene by introducing the concept of a spatial distribution of Gaussians (SDG), [29]. After affine motion compensation, a MAP decision criteria is used to label a pixel based on its intensity and spatial membership probabilities (both modeled as Gaussian *pdf*s). There are two primary points of interest in [29]. Firstly, the authors modeled the spatial position as a *single* Gaussian, negating the possibility of bimodal or multi-modal *spatial* probabilities, i.e. that a certain background element model may be expected to occur in more than one position. Although, not within the scope of their problem definition, this is, in fact, a definitive feature of a temporal texture. Analogous to the need for a mixture model to describe intensity distributions, unimodal distributions are limited in their ability to model

spatial uncertainty. 'Nonstationary' backgrounds have most recently been addressed by Pless *et al* [28] and Mittal *et al* [24]. Pless *et al* proposed several pixel-wise models based on the distributions of the image intensities and spatio-temporal derivatives. Mittal *et al* proposed an adaptive kernel density estimation scheme with a joint pixel-wise model of color (for a normalized color space), and optical flow at each pixel. Other notable pixel-wise detection schemes include [34], where topology free HMMs are described and several state splitting criteria are compared in context of background modeling, and [30], where a (practically) non-adaptive three-state HMM is used to model the background.

The second category of methods use region models of the background. In [35], Toyama *et al* proposed a three tiered algorithm that used region based (spatial) scene information in addition to per-pixel background model: region and frame level information served to verify pixel-level inferences. Another global method proposed by Oliver *et al* [26] used eigenspace decomposition to detect objects. For k input frames of size $N \times M$ a matrix $\mathbf{B}$ of size $k \times (NM)$ was formed by row-major vectorization of each frame and eigenvalue decomposition was applied to $\mathbf{C} = (\mathbf{B} - \mu)^{\mathbf{T}}(\mathbf{B} - \mu)$. The background was modeled by the eigenvectors corresponding to the η largest eigenvalues, $\mathbf{u_i}$, that encompass possible illuminations in the field of view (FOV). Thus, this approach is less sensitive to illumination. The foreground objects are detected by projecting the current image in the eigenspace and finding the difference between the reconstructed and actual images. The most recent region-based approaches are by Monnet *et al* [25], Zhong *et al* [40]. Monnet *et al* and Zhong *et al* simultaneously proposed models of image regions as an autoregressive moving average (ARMA) process, which is used to incrementally learn (using PCA) and then predict motion patterns in the scene.

The foremost assumption made in background modeling is the assumption of a station-

ary scene. However, this assumption is violated fairly regularly, through common real world phenomenon like swaying trees, water ripples, fountains, escalators etc. The local search proposed in [7], the SDG of [29], the time series models of [25], [40] and KDEs over color and optical flow in [24] are several formulations proposed for detection non-stationary backgrounds. While each method demonstrated degrees of success, the issue of spatial dependencies has not been addressed in a principled manner. In context of earlier work (in particular [24]), our approach falls under the category of methods that employ regional models of the background. We assert that useful correlation exists in the intensities of spatially proximal pixels and this correlation can be used to allow high levels of detection accuracy in the presence of general non-stationary phenomenon.

B. Proposed Formulation

The proposed work has three novel contributions. Firstly, the method proposed here provides a principled means of modeling the spatial dependencies of observed intensities. The model of image pixels as independent random variables, an assumption almost ubiquitous in background subtraction methods, is challenged and it is further asserted that there exists useful structure in the spatial proximity of pixels. This structure is exploited to sustain high levels of detection accuracy in the presence of nominal camera motion and dynamic textures. By using nonparametric density estimation methods over a joint domain-range representation, the background data is modeled as a single distribution and multi-modal spatial uncertainties can be directly handled. Secondly, unlike previous approaches, the foreground is explicitly modeled to augment the detection of objects without using tracking information. The criterion of temporal persistence is proposed for simultaneous use with the conventional criterion of background difference. Thirdly, instead of directly applying a threshold to membership probabilities, which implicitly assumes

independence of labels, we propose a MAP-MRF framework that competitively uses the foreground and background models for object detection, while enforcing spatial context in the process.

II. Object Detection

In this section we describe the novel representation of the background, the use of temporal persistence to pose object detection as a genuine binary classification problem, and the overall MAP-MRF decision framework. For an image of size $M \times N$, let $\mathcal{S}$ discretely and regularly index the image lattice, $\mathcal{S} = \{(i,j)|\ 1 \leq i \leq N, 1 \leq j \leq M\}$. In context of object detection in a stationary camera, the objective is to assign a binary label from the set $\mathcal{L} = \{\text{background}, \text{foreground}\}$ to each of the sites in $\mathcal{S}$.

A. Joint Domain-Range Background Model

If the primary source of spatial uncertainty of a pixel is image misalignment, a Gaussian density would be an adequate model since the corresponding point in the subsequent frame is equally likely to lie in any direction. However, in the presence of dynamic textures, cyclic motion, and non-stationary backgrounds in general, the 'correct' model of spatial uncertainty often has an arbitrary shape and may be bi-modal or multi-modal, but structure exists because by definition, the motion follows a certain repetitive pattern. Such arbitrarily structured data can be best analyzed using nonparametric methods since these methods make no underlying assumptions on the shape of the density. Non-parametric estimation methods operate on the principle that dense regions in a given feature space, populated by feature points from a class, correspond to the modes of the 'true' *pdf*. In this work, analysis is performed on a feature space where the p pixels are represented by $\mathbf{x}_i \in \mathbb{R}^5$, $i = 1, 2, \ldots p$. The feature vector, $\mathbf{x}$, is a joint domain-range representation,

where the space of the image lattice is the *domain*, (x, y) and some color space, for instance (r, g, b), is the *range*, [4]. Using this representation allows a *single* model of the entire background, $f_{R,G,B,X,Y}(r, g, b, x, y)$, rather than a collection of pixel-wise models. Pixel-wise models ignore the dependencies between proximal pixels and it is asserted here that these dependencies are important. The joint representation provides a direct means to model and exploit this dependency.

In order to build a background model, consider the situation at time t, before which all pixels, represented in 5-space, form the set $\psi_b = \{\mathbf{y_1}, \mathbf{y_2} \dots \mathbf{y_n}\}$ of the background. Given this sample set, at the observation of the frame at time t, the probability of each pixel-vector belonging to the background can be computed using the kernel density estimator ([27], [31]). The kernel density estimator is a nonparametric estimator and under appropriate conditions the estimate it produces is a valid probability itself. Thus, to find the probability that a candidate point, $\mathbf{x}$, belongs to the background, ψ_b, an estimate can be computed,

$$P(\mathbf{x}|\psi_\mathbf{b}) = \mathbf{n^{-1}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \varphi_\mathbf{H}\Big(\mathbf{x} - \mathbf{y_i}\Big), \tag{1}$$

where $\mathbf{H}$ is a symmetric positive definite $d \times d$ bandwidth matrix, and

$$\varphi_\mathbf{H}(\mathbf{x}) = |\mathbf{H}|^{\mathbf{-1/2}} \varphi(\mathbf{H}^{\mathbf{-1/2}}\mathbf{x}), \tag{2}$$

where φ is a d-variate kernel function usually satisfying $\int \varphi(\mathbf{x})\mathbf{dx} = \mathbf{1}$, $\varphi(\mathbf{x}) = \varphi(-\mathbf{x})$, $\int \mathbf{x}\varphi(\mathbf{x})\mathbf{dx} = \mathbf{0}$, $\int \mathbf{x}\mathbf{x}^\mathbf{T}\varphi(\mathbf{x})\mathbf{dx} = \mathbf{I_d}$ and is also usually compactly supported. The d-variate Gaussian density is a common choice as the kernel φ,

$$\varphi_\mathbf{H}^{(\mathcal{N})}(\mathbf{x}) = |\mathbf{H}|^{\mathbf{-1/2}} (\mathbf{2}\pi)^{\mathbf{-d/2}} \exp\Big(-\frac{\mathbf{1}}{\mathbf{2}}\mathbf{x}^\mathbf{T}\mathbf{H}^{\mathbf{-1}}\mathbf{x}\Big). \tag{3}$$

It is stressed here, that using a Gaussian kernel does not make any assumption on the scatter of data in the feature space. The kernel function only defines the effective region of

influence of each data point while computing the final probability estimate. Any function that satisfies the constraints specified after Equation 2, i.e. a valid pdf, symmetric, zero-mean, with identity covariance, can be used as a kernel. There are other functions that are commonly used, some popular alternatives to the Gaussian kernel are the Epanechnikov kernel, the Triangular kernel, the Bi-weight kernel and the Uniform kernel, each with their merits and demerits (see [38] for more details).

Within the joint domain-range feature space, the kernel density estimator explicitly models spatial dependencies, without running into difficulties of parametric modeling. Furthermore, since it is well known that the rgb axes are correlated, it is worth noting that kernel density estimation also accounts for this correlation. The result is a single model of the background.

Lastly, in order to ensure that the algorithm remains adaptive to slower changes (such as illumination change or relocation) a sliding window of length ρ_b frames is maintained. This parameter corresponds to the learning rate of the system.

A.1 Bandwidth Estimation

Asymptotically, the selected bandwidth $\mathbf{H}$ does not affect the kernel density estimate but in practice sample sizes are limited. Too small a choice of $\mathbf{H}$ and the estimate begins to show spurious features, too large a choice of $\mathbf{H}$ leads to an over-smoothed estimate, losing important structural features like multi-modality. In general, rules for choosing bandwidths are based on balancing bias and variance globally. Theoretically, the ideal or optimal $\mathbf{H}$ can be found by minimizing the mean-squared error,

$$MSE\{\hat{f}_{\mathbf{H}}(\mathbf{x})\} = \mathbf{E}\{[\hat{\mathbf{f}}_{\mathbf{H}}(\mathbf{x}) - \mathbf{f}_{\mathbf{H}}(\mathbf{x})]^{\mathbf{2}}\}, \tag{4}$$

where $\hat{f}$ is the estimated density and f is the true density. Evidently, the optimal value of

$\mathbf{H}$ is data dependent since the MSE value depends on $\mathbf{x}$. However, in practice, one does not have access to the true density function which is required to estimate the optimal bandwidth. Instead, a fairly large number of heuristic approaches have been proposed for finding $\mathbf{H}$, a survey is provided in [36].

Adaptive estimators have been shown to considerably outperform (in terms of the mean squared error) the fixed bandwidth estimator, particularly in higher dimensional spaces, [32]. In general two formulations of adaptive or variable bandwidth estimators have been considered [19]. The first varies the bandwidth with the estimation point and is called the balloon estimator, given by,

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{\mathbf{H}(\mathbf{x})}(\mathbf{x} - \mathbf{x_i})), \tag{5}$$

where $\mathbf{H}(\mathbf{x})$ is the bandwidth matrix at $\mathbf{x}$. The second approach, called the sample-point estimator, varies the bandwidth matrix depending on the sample point,

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{\mathbf{H}(\mathbf{x_i})}(\mathbf{x} - \mathbf{x_i})). \tag{6}$$

where $\mathbf{H}(\mathbf{x_i})$ is the bandwidth matrix at $\mathbf{x_i}$. However, developing variable bandwidth schemes for kernel density estimation is still research in progress, both in terms of theoretical understanding and in terms of practical algorithms, [32].

In the given application, the sample size is large, and although it populates a 5 dimensional feature space, the estimate was found to be reasonably robust to the selection of bandwidth. Furthermore, choosing an optimal bandwidth in the MSE sense is usually highly computationally expensive. Thus, the balance between accuracy required (for matting, object recognition or action recognition) and computational speed (for real-time surveillance systems) is application specific. To reduce the computational load, the Binned

kernel density estimator provides a practical means of dramatically increasing computational speeds while closely approximating the kernel density estimate of Equation 1, ([38], Appendix D). With appropriate binning rules and kernel functions the accuracy of the the Binned KDE is shown to approximate the kernel density estimate in [13]. Binned versions of the adaptive kernel density estimate have also been provided in [32]. To further reduce computation, the bandwidth matrix $\mathbf{H}$ is usually either assumed to be of the form $\mathbf{H} = \mathbf{h^2 I}$ or $\mathbf{H} = \mathbf{diag}(\mathbf{h_1^2}, \mathbf{h_2^2}, \ldots \mathbf{h_d^2})$. Thus, rather than selecting a fully parameterized bandwidth matrix, only two parameters can be defined, one for the variance in the spatial dimensions (x, y) and and one for the color channels, reducing computational load.

B. Modeling the Foreground

The intensity difference of interesting objects from the background has been, by far, the most widely used criterion for object detection. In this paper, *temporal persistence* is proposed as a property of real foreground objects, i.e. *interesting objects tend to remain in the same spatial vicinity and tend to maintain consistent colors from frame to frame.* The joint representation used here allows competitive classification between the foreground and background. To that end, models for both the background and the foreground are maintained. An appealing feature of this representation is that the foreground model can be constructed in a consistent fashion with the background model: a joint domain-range non-parametric density $\psi_f = \{\mathbf{z_1}, \mathbf{z_2} \ldots \mathbf{z_m}\}$. Just as there was a learning rate parameter ρ_b for the background model, a parameter ρ_f is defined for the foreground frames. However, since the foreground changes far more rapidly than the background, the learning rate of the foreground is typically much higher than that of the background.

At any time instant the probability of observing a foreground pixel at any location (i, j) of any color is uniform. Then, once a foreground region is been detected at time t, there

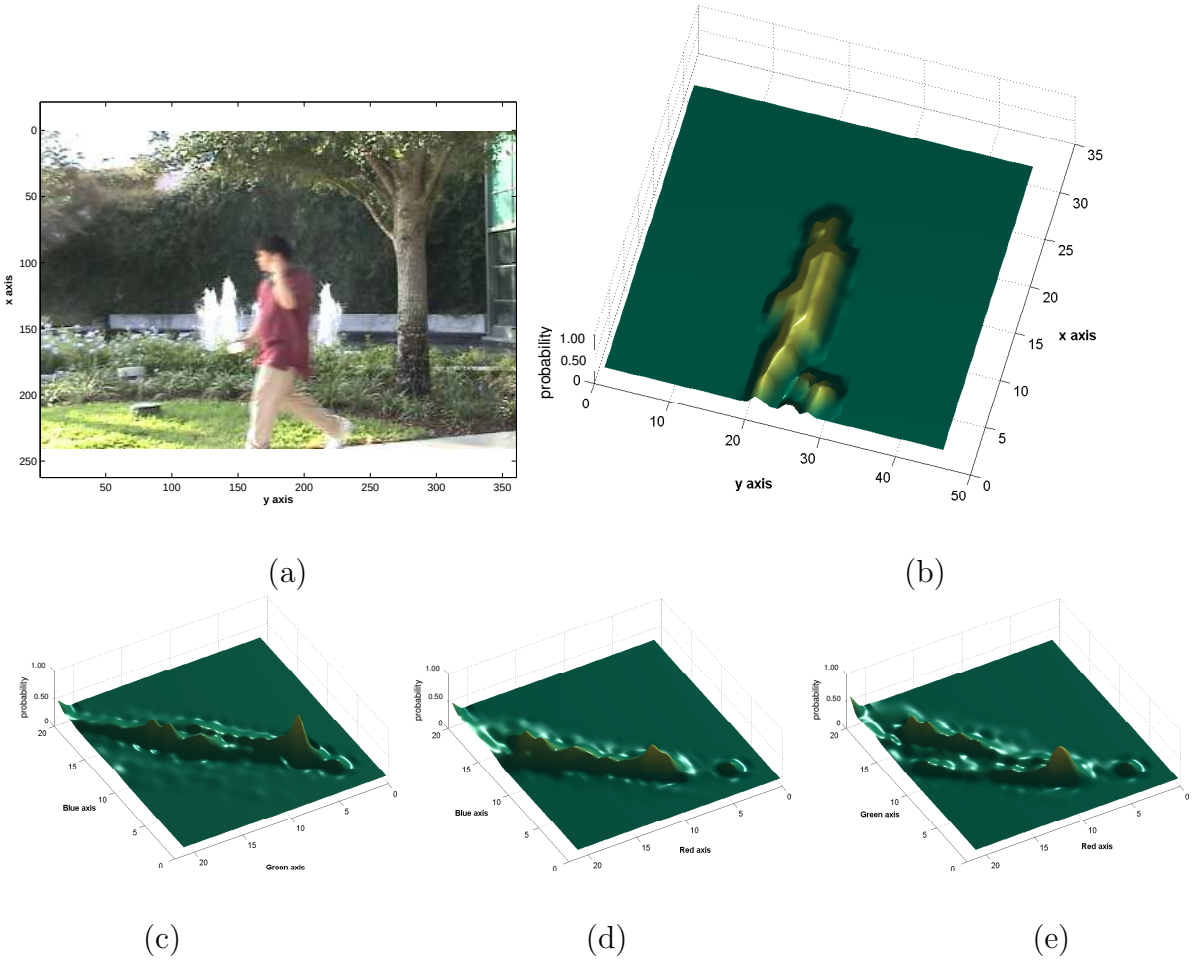


Fig. 2. Foreground Modeling. Using kernel density estimates on a model built from recent frames, the foreground can be detected in subsequent frames using the property of temporal persistence, (a) Current Frame (b) the X, Y-marginal, $f_{X,Y}(x, y)$. High membership probabilities are seen in regions where foreground in the current frame matches the recently detected foreground. The non-parametric nature of the model allows the arbitrary shape of the foreground to be captured accurately (c) the B, G-marginal, $f_{B,G}(b, g)$ (d) the B, R-marginal, $f_{B,R}(b, r)$ (e) the G, R-marginal, $f_{G,R}(g, r)$.

is an increased probability of observing a foreground region at time $t + 1$ in the same proximity with a similar color distribution. Thus, foreground probability is expressed as a mixture of a uniform function and the kernel density function,

$$P(\mathbf{x}|\psi_{\mathbf{f}}) = \alpha\gamma + (\mathbf{1} - \alpha)\mathbf{m^{-1}} \sum_{\mathbf{i=1}}^{\mathbf{m}} \varphi_{\mathbf{H}}\Big(\mathbf{x} - \mathbf{z_i}\Big), \tag{7}$$

where $\alpha \ll 1$ is the mixture weight, and γ is a random variable with uniform probability,

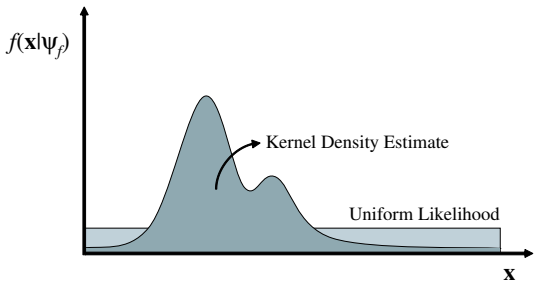


Fig. 3. Foreground likelihood function. The foreground likelihood estimate is a mixture of the kernel density estimate and a uniform likelihood across the 5-space of features. This figure shows a conceptualization as a 1-D function.

that is $\gamma_{R,G,B,X,Y}(r,g,b,x,y) = \frac{1}{R\times G\times B\times M\times N}$, where $0 \leq r \leq R,\ 0 \leq g \leq G,\ 0 \leq b \leq B,\ 0 \leq x \leq M,\ 0 \leq y \leq N$. This mixture is illustrated in Figure 3. If an object is detected in the preceding frame, the probability of observing the colors of that object in the same proximity increases according to the second term in Equation 7. Therefore, as objects of interest are detected (the detection method will be explained presently) all pixels that are classified as 'interesting' are used to update the foreground model ψ_f. In this way, simultaneous models are maintained of both the background and the foreground, which are then used competitively to estimate interesting regions. Finally, to allow objects to become part of the background (e.g. a car having been parked or new construction in an environment), all pixels are used to update ψ_b. Figures 2 shows plots of some marginals of the foreground model.

At this point, whether a pixel vector $\mathbf{x}$ is 'interesting' or not can be competitively estimated using a simple *likelihood ratio classifier* (or a Parzen Classifier since likelihoods are computed using Parzen density estimates, [10]),

$$\tau = -\ln \frac{P(\mathbf{x}|\psi_\mathbf{b})}{P(\mathbf{x}|\psi_\mathbf{f})} = -\ln \frac{n^{-1}\sum_{i=1}^{n}\varphi_\mathbf{H}\Big(\mathbf{x}-\mathbf{y_i}\Big)}{\alpha\gamma + (1-\alpha)m^{-1}\sum_{i=1}^{m}\varphi_\mathbf{H}\Big(\mathbf{x}-\mathbf{z_i}\Big)} \tag{8}$$

Thus the classifier δ is,

$$\delta(\mathbf{x}) = \begin{cases} -1 & \text{if } -\ln \frac{P(\mathbf{x}|\psi_{\mathbf{b}})}{P(\mathbf{x}|\psi_{\mathbf{f}})} > \kappa \\ 1 & \text{otherwise} \end{cases}$$

where κ is a threshold which balances the trade-off between sensitivity to change and robustness to noise. The utility in using the foreground model for detection can be clearly seen in Figure 4. Figure 4(e) shows the likelihood values based only on the background model and Figure 4(f) shows the likelihood ratio based on both the foreground and the background models. In both histograms, two processes can be roughly discerned, a major one corresponding to the background pixels and a minor one corresponding to the foreground pixels. The variance *between* the clusters increases with the use of the foreground model. Visually, the areas corresponding to the tires of the cars are positively affected, in particular. The final detection for this frame is shown in Figure 8(c). Evidently, the higher the likelihood of belonging to the foreground, the lower the overall likelihood ratio. However, as is described next, instead of using only likelihoods, prior information of neighborhood spatial context is enforced in a MAP-MRF framework. This removes the need to specify the arbitrary parameter κ.

C. Spatial Context: Estimation using a MAP-MRF Framework

The inherent spatial coherency of objects in the real world is often applied in a post-processing step, in the form of morphological operators like erosion and dilation, by using a median filter or by neglecting connected components containing only a few pixels, [33]. Furthermore, directly applying a threshold to membership probabilities implies conditional independence of labels, i.e. $P(\ell_i|\ell_j) = P(\ell_i)$, where $i \neq j$, and ℓ_i is the label of pixel i. We assert that such conditional independence rarely exists between proximal sites. Instead of applying such ad-hoc heuristics, Markov Random Fields provide a math-

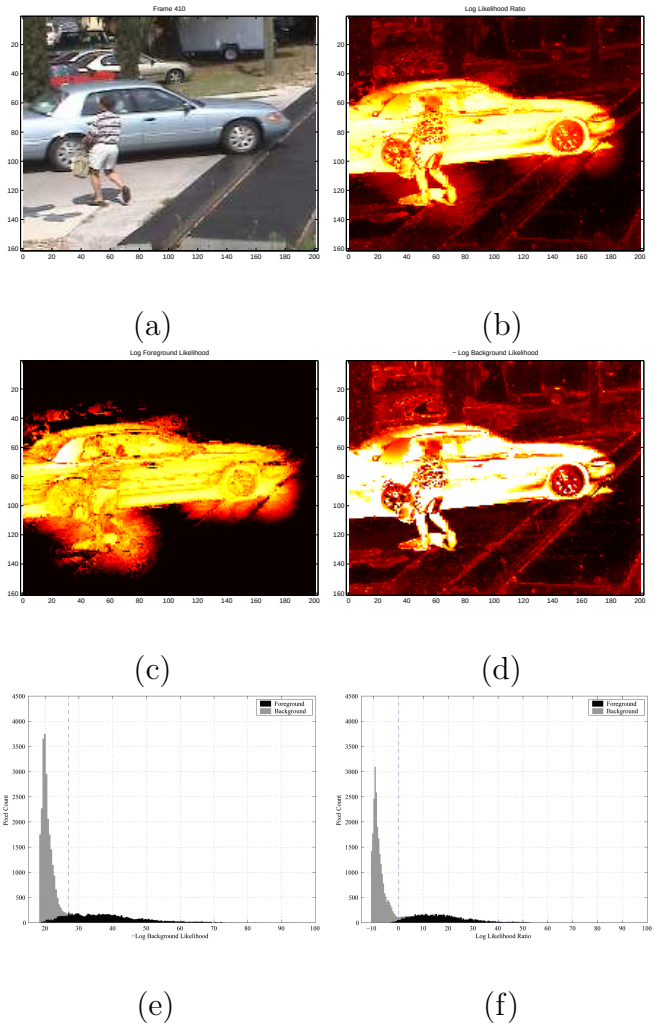
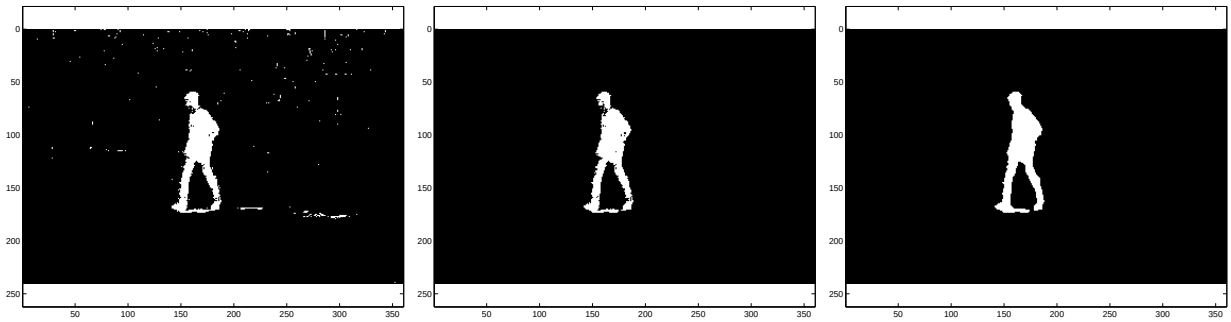


Fig. 4. Improvement in discrimination using temporal persistence. Whiter values correspond to higher likelihoods of foreground membership. (a) Video Frame 410 of the Nominal Motion Sequence (b) Log-Likelihood Ratio values obtained using Equation 8. (c) Foreground likelihood map. (d) Background negative log-likelihood map. (e) Histogrammed negative log-likelihood values for background membership. The dotted line represents the 'natural' threshold for the background likelihood, i.e. $\log(\gamma)$. (f) Histogrammed log-likelihood ratio values. Clearly the variance *between* clusters is decidedly enhanced. The dotted line represents the 'natural' threshold for the log-likelihood ratio, i.e. zero.

ematical foundation to make a global inference using local information. While in some instances the morphological operators may do as well as the MRF for removing residual mis-detections at a reduced computational cost, there are two central reasons for using the MRF:



(a) (b) (c)

Fig. 5. Three possible detection strategies. (a) Detection by thresholding using only the background model of Equation 1. Noise can cause several spurious detections. (b) Detection by thresholding the Likelihood Ratio of Equation 8. Since some spurious detections do not persist in time, false positives are reduced using the foreground model. (c) Detection using MAP-MRF estimation, 13. All spurious detections are removed and false negative within the detected object are also removed as a result of their spatial context.

1. By selecting an edge-preserving MRF, the resulting smoothing will respect the object boundaries.

2. As will be seen, the formulation of the problem using the MRF introduces regularity into the final energy function that allows for the optimal partition of the frame (through computation of the minimum cut), without the need to pre-specify the parameter κ.

3. The MRF prior is precisely the constraint of spatial context we wish to impose on $\mathcal{L}$. For the MRF, the set of neighbors, $\mathcal{N}$, is defined as the set of sites within a radius $r \in \mathbb{R}$ from site $\mathbf{i} = (\mathbf{i}, \mathbf{j})$,

$$\mathcal{N}_{\mathbf{i}} = \{\mathbf{u} \in \mathcal{S} |\ \mathit{distance}(\mathbf{i}, \mathbf{u}) \leq \mathbf{r}, \mathbf{i} \neq \mathbf{u}\}, \tag{9}$$

where $\mathit{distance}(\mathbf{a}, \mathbf{b})$ denotes the Euclidean distance between the pixel locations $\mathbf{a}$ and $\mathbf{b}$. The 4-neighborhood (used in this paper) and 8-neighborhood cliques are two commonly used neighborhoods. The pixels $\hat{\mathbf{x}} = \{\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_p}\}$ are conditionally independent given $\mathcal{L}$, with conditional density functions $f(\mathbf{x_i}|\ell_\mathbf{i})$. Thus, since each $\mathbf{x_i}$ is dependant on $\mathcal{L}$ only

through ℓ_i, the likelihood function may be written as,

$$l(\hat{\mathbf{x}}|\mathcal{L}) = \prod_{\rangle=\infty}^{\surd} \{(\mathbf{x_i}|\ell_\mathbf{i}) = \prod_{\mathbf{i=1}}^{\mathbf{p}} \mathbf{f}(\mathbf{x_i}|\psi_\mathbf{f})^{\ell_\mathbf{i}} \mathbf{f}(\mathbf{x_i}|\psi_\mathbf{b})^{\mathbf{1}-\ell_\mathbf{i}}. \tag{10}$$

Spatial context is enforced in the decision through a pairwise interaction MRF prior. We use the Ising Model for its discontinuity preserving properties,

$$p(\mathcal{L}) \propto \exp\Big(\sum_{i=1}^{p}\sum_{j=1}^{p} \lambda\Big(\ell_i\ell_j + (1-\ell_i)(1-\ell_j)\Big)\Big), \tag{11}$$

where λ is a positive constant and $i \neq j$ are neighbors. By Bayes Law, the posterior, $p(\mathcal{L}|\hat{\mathbf{x}})$, is then equivalent to

$$p(\mathcal{L}|\hat{\mathbf{x}}) = \frac{\mathbf{p}(\hat{\mathbf{x}}|\mathcal{L})\mathbf{p}(\mathcal{L})}{\mathbf{p}(\hat{\mathbf{x}})} = \frac{\Big(\prod_{i=1}^{p} f(\mathbf{x_i}|\psi_\mathbf{f})^{\ell_\mathbf{i}} \mathbf{f}(\mathbf{x_i}|\psi_\mathbf{b})^{\mathbf{1}-\ell_\mathbf{i}}\Big)\mathbf{p}(\mathcal{L})}{p(\hat{\mathbf{x}})}. \tag{12}$$

Ignoring $p(\hat{\mathbf{x}})$ and other constant terms, the log-posterior, $\ln p(\mathcal{L}|\hat{\mathbf{x}})$, is then equivalent to,

$$L(\mathcal{L}|\hat{\mathbf{x}}) = \sum_{\mathbf{i=1}}^{\mathbf{p}} \ln\left(\frac{\mathbf{f}(\mathbf{x_i}|\psi_\mathbf{f})}{\mathbf{f}(\mathbf{x_i}|\psi_\mathbf{b})}\right)\ell_\mathbf{i} + \sum_{i=1}^{p}\sum_{j=1}^{p} \lambda\Big(\ell_i\ell_j + (1-\ell_i)(1-\ell_j)\Big). \tag{13}$$

The MAP estimate is the binary image that maximizes L and since there are 2^{NM} possible configurations of $\mathcal{L}$ an exhaustive search is usually infeasible. In fact, it is known that minimizing discontinuity-preserving energy functions in general is NP-Hard, [2]. Although, various strategies have been proposed to minimize such functions, e.g. Iterated Condition Modes [1] or Simulated Annealing [11], the solutions are usually computationally expensive to obtain and of poor quality. Fortunately, since L belongs to the $\mathcal{F}^2$ class of energy functions, defined in [22] as a sum of function of up to two binary variables at a time,

$$E(x_1, \ldots x_n) = \sum_i E^i(x_i) + \sum_{i,j} E^{(i,j)}(x_i, x_j), \tag{14}$$

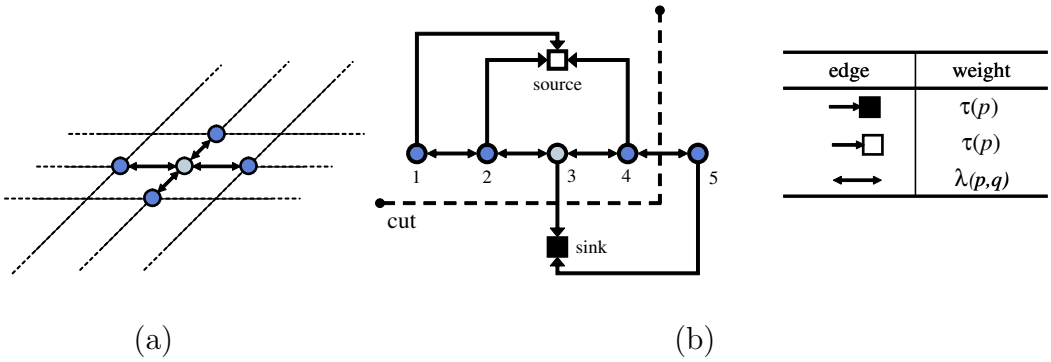


Fig. 6. A 4-neighborhood system. Each pixel location corresponds to a node in the graph, connected by a directed edge to the source and the sink, and by an undirected edge to it's four neighbors. For purposes of clarity the edges between node 3 and nodes 5 and 1 have been omitted in (b).

and since it satisfies the regularity condition of the so-called $\mathcal{F}^2$ theorem, efficient algorithms exist for the optimization of L by finding the minimum cut of a capacitated graph, [12], [22], described next.

Algorithm

Initialize ψ_b using 1^{st} frame, $\psi_f = \emptyset$. At frame t, for each pixel,

Detection Step

1. Find $P(\mathbf{x_i}|\psi_\mathbf{f})$ (Eq. 7) and $P(\mathbf{x_i}|\psi_\mathbf{b})$ (Eq. 1) and compute the Likelihood Ratio τ (Eq. 8).
2. Construct the graph to minimize Equation 13.

Model Update Step

1. Append all pixels detected as foreground to the foreground model ψ_f.
2. Remove all pixels in ψ_f from ρ_f frames ago.
3. Append all pixels of the image to the background model ψ_b.
4. Remove all pixels in ψ_b from ρ_b frames ago.

Fig. 7. Object Detection Algorithm

To maximize the energy function (Equation 13), we construct a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ with a 4-neighborhood system $\mathcal{N}$ as shown in Figure 6. In the graph, there are two distinct terminals s and t, the sink and the source, and n nodes corresponding to each

image pixel location, thus $\mathcal{V} = \{v_1, v_2, \cdots, v_n, s, t\}$. A solution is a two-set *partition*, $\mathcal{U} = \{s\} \cup \{i|\ell_i = 1\}$ and $\mathcal{W} = \{t\} \cup \{i|\ell_i = 0\}$. The graph construction is as described in [12], with a directed edge (s, i) from s to node i with a weight $w_{(s,i)} = \tau_i$ (the log-likelihood ratio), if $\tau_i > 0$, otherwise a directed edge (i, t) is added between node i and the sink t with a weight $w_{(i,t)} = -\tau_i$. For the second term in Equation 13, undirected edges of weight $w_{(i,j)} = \lambda$ are added if the corresponding pixels are neighbors as defined in $\mathcal{N}$ (in our case if j is within the 4-neighborhood clique of i) . The capacity of the graph is $C(\mathcal{L}) = \sum_i \sum_j w_{(i,j)}$, and a cut defined as the set of edges with a vertex in $\mathcal{U}$ and a vertex in $\mathcal{W}$. As shown in [8], the minimum cut corresponds to the maximum flow, thus maximizing $L(\mathcal{L}|\hat{\mathbf{x}})$ is equivalent to finding the minimum cut. The minimum cut of the graph can be computed through a variety of approaches, the Ford-Fulkerson algorithm or a faster version proposed in [12]. The configuration found thus corresponds to an optimal estimate of $\mathcal{L}$. The complete algorithm is described in Figure 7.

III. Results and Discussion

The algorithm was tested on a variety of sequences in the presence of nominal camera motion, dynamic textures, and cyclic motion. On a 3.06 GHz Intel Pentium 4 processor with 1 GB RAM, an optimized implementation of the proposed approach can process about 11 fps for a frame size of 240 $\times$ 360. The sequences were all taken with a COTS camera (the Sony DCR-TRV 740). Comparative results for the mixture of Gaussians method have also been shown. For all the results the bandwidth matrix $\mathbf{H}$ was parameterized as a diagonal matrix with three equal variances pertaining to the range (color), represented by h_r and two equal variances pertaining to the domain, represented by h_d. The values used in all experiments were $(h_r, h_d) = (16, 25)$.

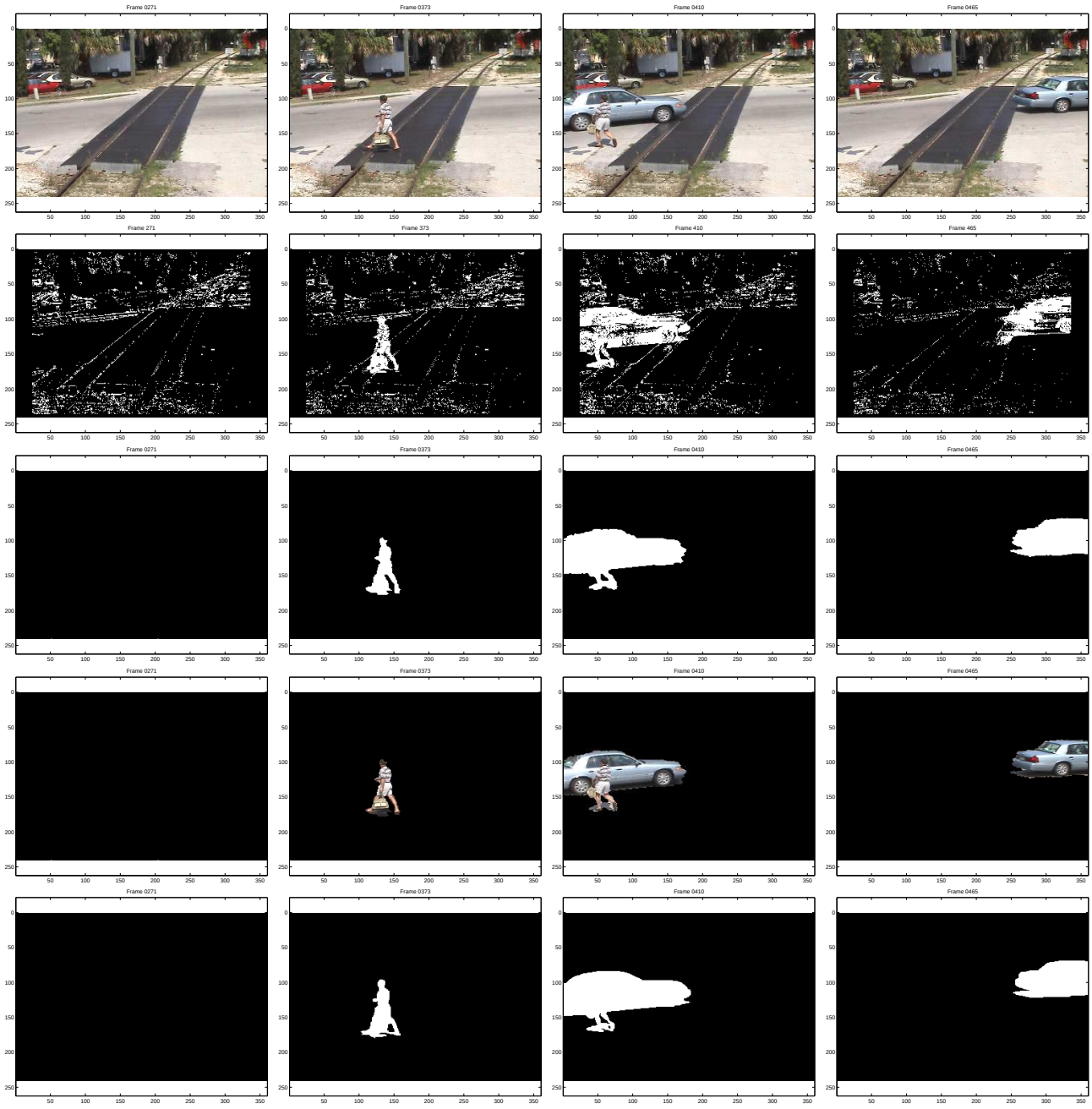
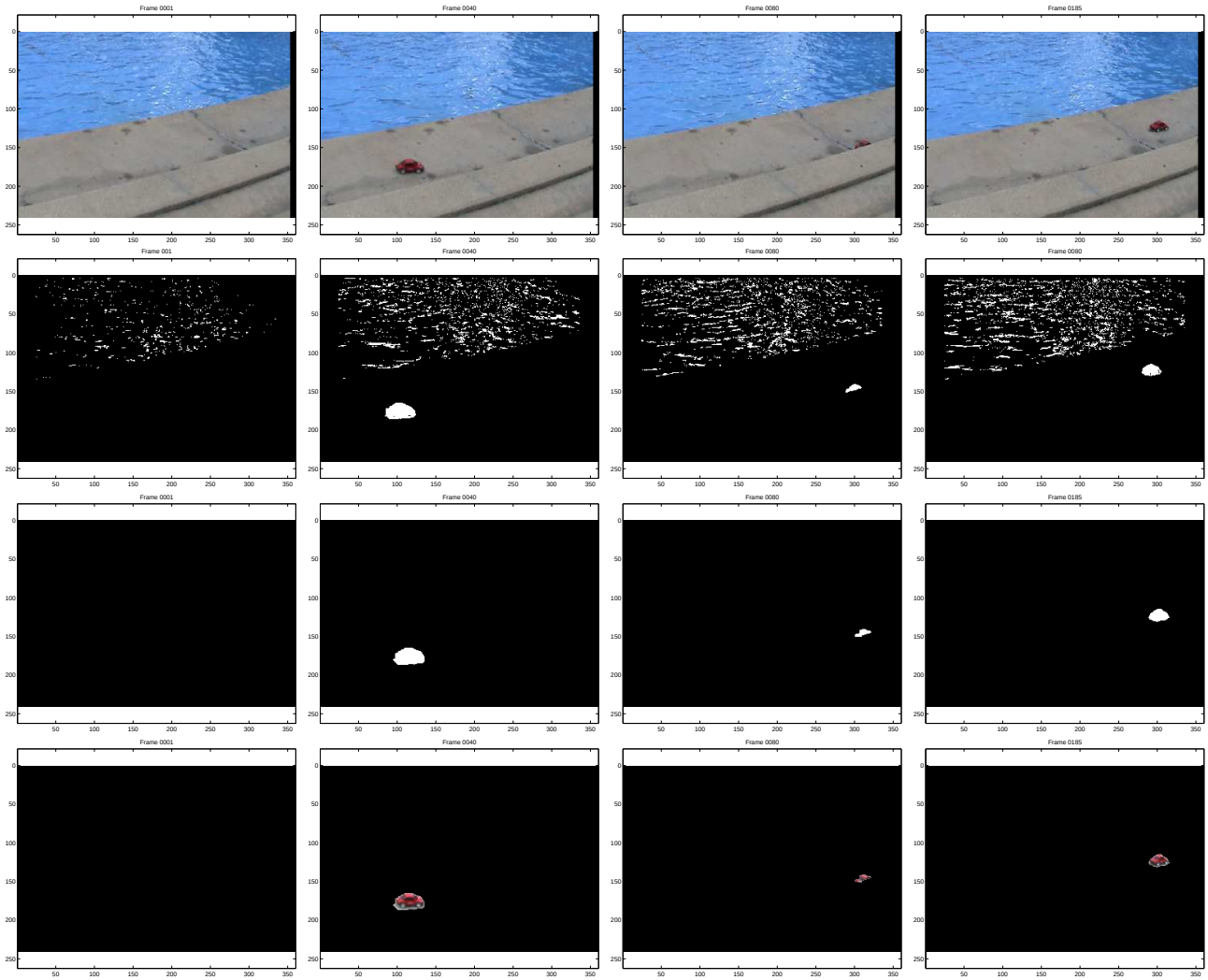
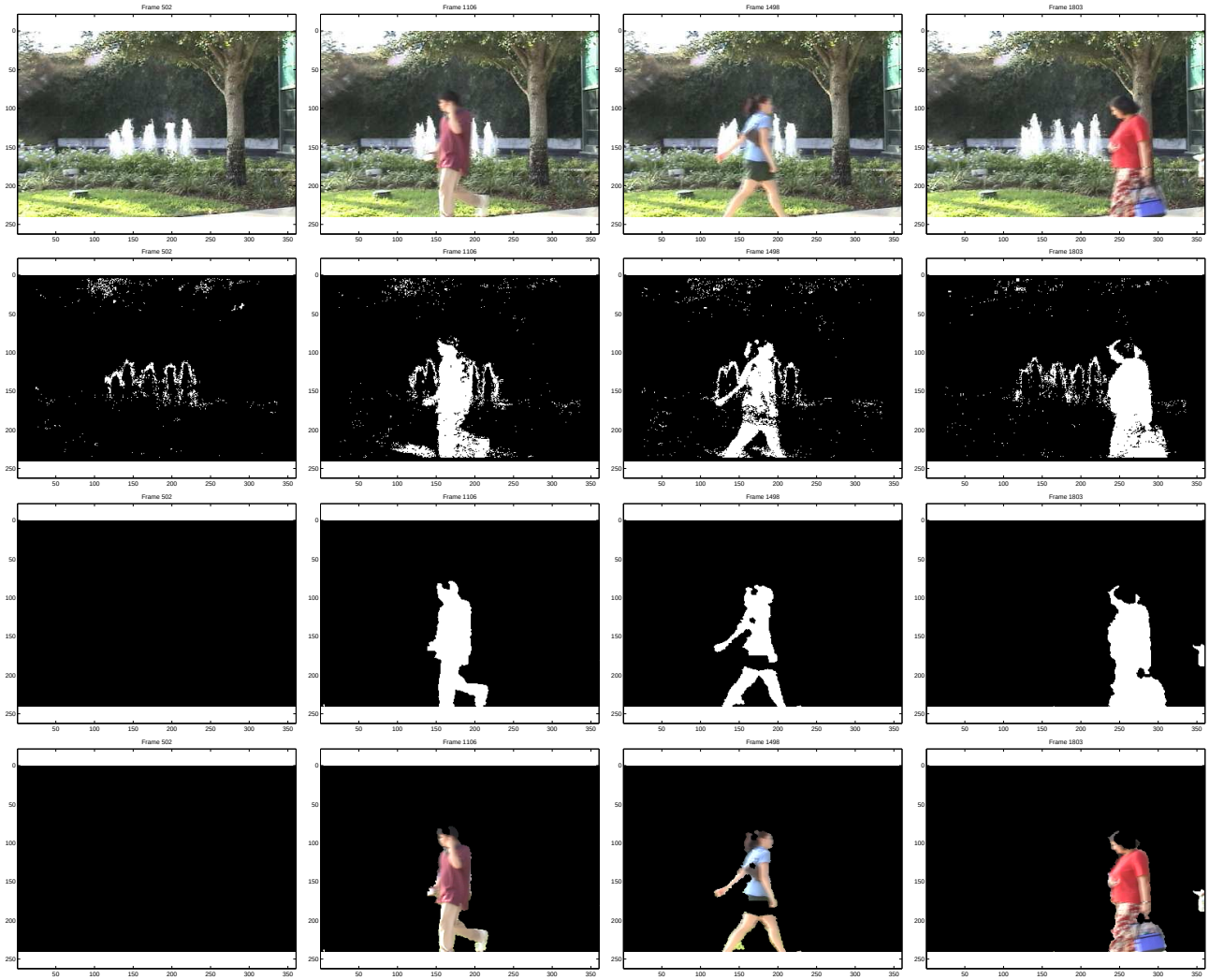


Fig. 8. Background Subtraction in a nominally moving camera (motion is an average of 12 pixels). The top row are the original images, the second row are the results obtained by using a 5-component, Mixture of Gaussians method, and the third row results obtained by the proposed method. The fourth row is the masked original image. The fifth row is the manual segmentation. Morphological operators were not used in the results.



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

Fig. 9. Poolside sequence. The water in this sequence shimmers and ripples causing false positive in conventional detection algorithms, as a remote controlled car passes on the side. The top row are the original images, the second row are the results obtained by using a 5-component, Mixture of Gaussians method, and the third row are the results obtained by the proposed method. The fourth row is the masked original image. Morphological operators were not used in the results.



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

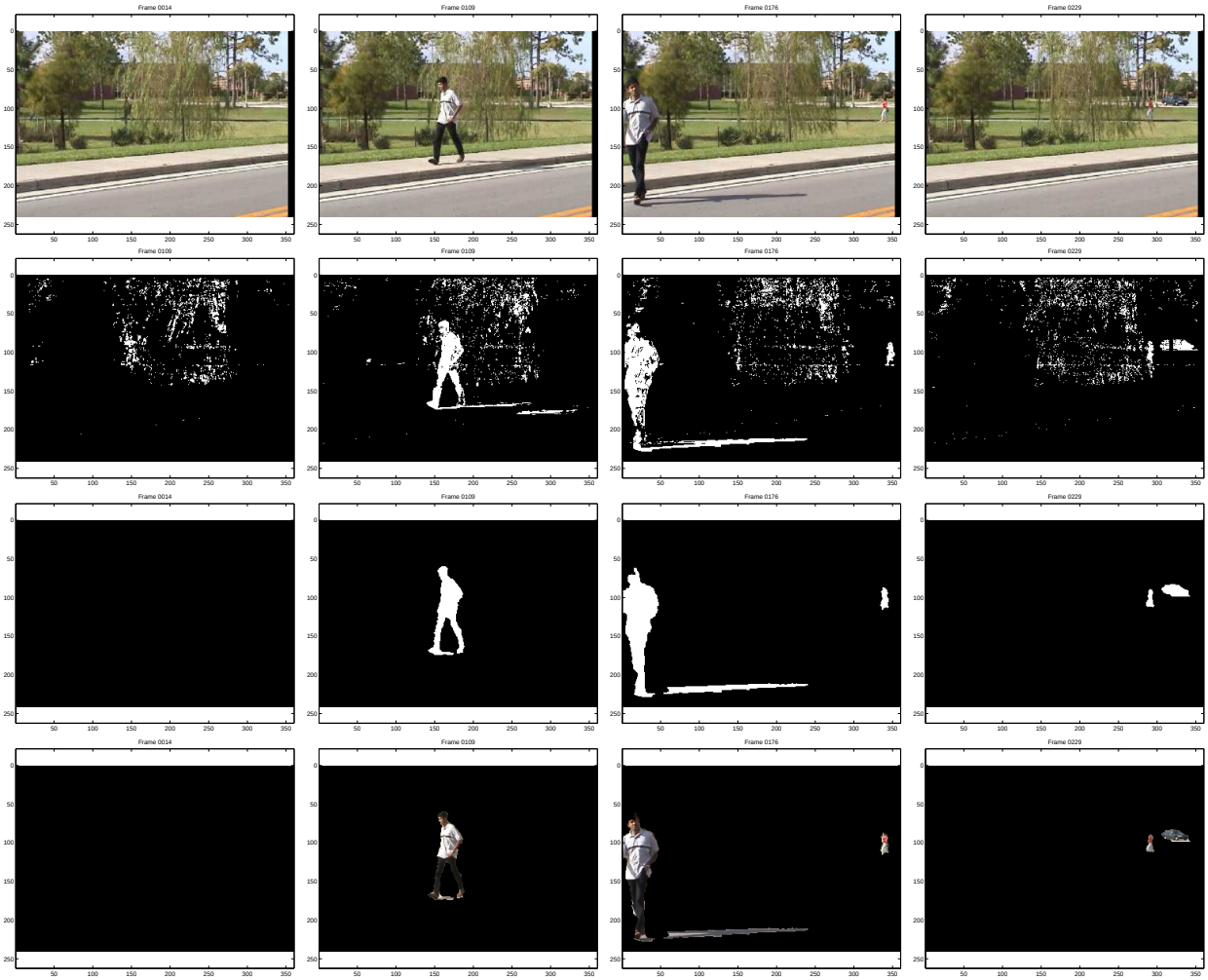
Fig. 10. Fountain Sequence. Background Subtraction in the presence of dynamic textures. There are three sources of nonstationarity: (1) The tree branches oscillate (2) The fountains (3) The shadow of the tree on the grass below. The top row are the original images, the second row are the results obtained by using a 5-component, Mixture of Gaussians method, and the third row results obtained by the proposed method. The fourth row is the masked original image. Morphological operators were not used in the results.

(a) (b) (c)

Fig. 11. Three more examples of detection in the presence of dynamic backgrounds. (a) The lake-side water is the source of dynamism in the background. The contour outlines the detected foreground region. (b) The periodic motion of the ceiling fans is ignored during detection. (c) A bottle floats on the oscillating sea, in the presence of rain.

A. Qualitative Analysis

Qualitative results on seven sequences of dynamic scenes are presented in this section. The first sequence that was tested involved a camera mounted on a tall tripod. The wind caused the tripod to sway back and forth causing nominal motion of the camera. Figure 8 shows the results obtained by the proposed algorithm. The first row are the recorded images, the second row shows the detected foreground as proposed in [33], and it is evident that the nominal motion of the camera causes substantial degradation in performance, despite a 5-component mixture model and a relatively high learning rate of 0.05. The third row shows the foreground detected using the proposed approach. It is stressed that *no* morphological operators like erosion / dilation or median filters were used in the presentation of these results. Manually segmented foreground regions are shown in the bottom row. This sequence exemplifies a set of phenomenon, including global motion caused by vibrations, global motion in static hand-held cameras, and misalignment in the registration of mosaics. Quantitative experimentation has been performed on this



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

Fig. 12. Swaying trees sequence. A weeping willow sways in the presence of a strong breeze. The top row shows the original images, the second row are the results obtained by using the mixture of Gaussians method, and the third row are the results obtained by the proposed method. The fourth row is the masked original image. Morphological operators were not used in the results.

sequence and is reported subsequently.

Figures 9, 10, and 12 show results on scenes with dynamic textures. In Figure 9, a red remote controlled car moves in a scene with a backdrop of a shimmering and rippling pool. Since dynamic textures like the water do not repeat exactly, pixel-wise methods, like the mixture of Gaussians approach, handle the dynamic texture of the pool poorly, regularly

producing false positives. On the other hand, the proposed approach handled this dynamic texture immediately, while detecting the moving car accurately as well. Figure 10 shows results on a particularly challenging outdoor sequence, with three sources of dynamic motion: (1) The fountain, (2) the tree branches above, and (3) the shadow of the trees branches on the grass below. The proposed approach disregarded each of the dynamic phenomena and instead detected the objects of interest. In Figure 12, results are shown on sequence where a weeping willow is swaying in a strong breeze. There were two typical paths in this sequence, one closer to the camera, and another one farther back, behind the tree. Including invariance to the dynamic behavior of the background, both the larger objects closer by and the smaller foreground objects farther back were detected as shown in Figure 12(c) and (d).

Figure 11(a) shows detection in the presence of period motion, a number of ceiling fans. Despite a high degree of motion, the individual is detected accurately. Figure 11(b) shows detection with the backdrop of a lake, and and 11(c) shows detection in the presence of substantial wave motion and rain. In each of the results of 11, the contour outlines the detected region, demonstrating accurate detection.

B. Quantitative Analysis

We performed quantitative analysis at both the pixel-level and object-level. For the first experiment, we manually segmented a 500-frame sequence (as seen in Figure 8) into foreground and background regions. In the sequence, the scene is empty for the first 276 frames, after which two objects (first a person and then a car) move across the field of view. The sequence contained an average nominal motion of approximately 14.66 pixels. Figure 13(a) shows the number of pixels detected in selected frames by the mixture of Gaussians method at various values of the learning parameter and the ground truth. The periodicity

apparent in the detection by the mixture of Gaussians method is caused by the periodicity of the camera motion. The initial periodicity in the ground truth is caused by the periodic self-occlusion of the walking person and the subsequent peak is caused by the later entry and then exit of the car. In Figure 13(b) the corresponding plot at each level of the proposed approach is shown. The threshold for the detection using only the background model was chosen as $\log(\gamma)$ (see Equation 7), which was equal to -27.9905. In addition to illustrating the contribution of background model to the over-all result, the performance at this level is also relevant because, in the absence of any previously detected foreground, the system essentially uses only the background model for detection. For the log-likelihood ratio, the obvious value for κ (see Equation 8) is zero, since this means the background is less likely than the foreground. Clearly, the results reflect the invariance at each level of the proposed approach to mis-detections caused by the nominal camera motion. The per-frame detection rates are shown in Figure 14 and Figure 15 in terms of precision and recall, where $\text{Precision} = \frac{\text{\# of true positives detected}}{\text{total \# of positives detected}}$ and $\text{Recall} = \frac{\text{\# of true positives detected}}{\text{total \# of true positives}}$. The detection accuracy both in terms of recall and precision is consistently higher than the mixture of Gaussians approach. Several different parameter configurations were tested for the mixture of Gaussians approach and the results are shown for three different learning parameters. The few false positives and false negatives that were detected by the proposed approach were invariably at the edges of true objects, where factors such as pixel sampling affected the results.

Next, to evaluate detection at the object level (detecting whether an object is present or not), we evaluated five sequences, each (approximately) an hour long. The sequences tested included an extended sequence of Figure 8, a sequence containing trees swaying in the wind, a sequence of ducks swimming on a pond, and two surveillance videos.

	Objects	Det.	Mis-Det.	Det. %	Mis-Det. %
Seq. 1	84	84	0	100.00%	0.00%
Seq. 2	115	114	1	99.13%	0.87%
Seq. 3	161	161	0	100.00%	0.00%
Seq. 4	94	94	0	100.00%	0.00%
Seq. 5	170	169	2	99.41%	1.18%

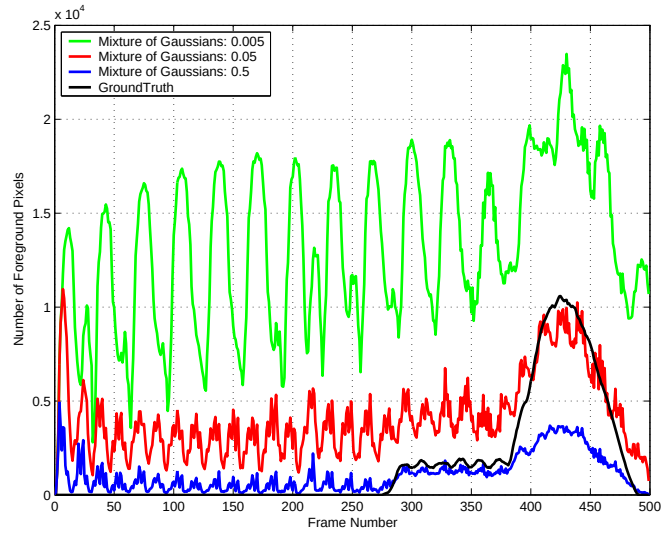
TABLE I

Object level detection rates. Object detection and mis-detection rates for 5 sequences (each 1 hour long).

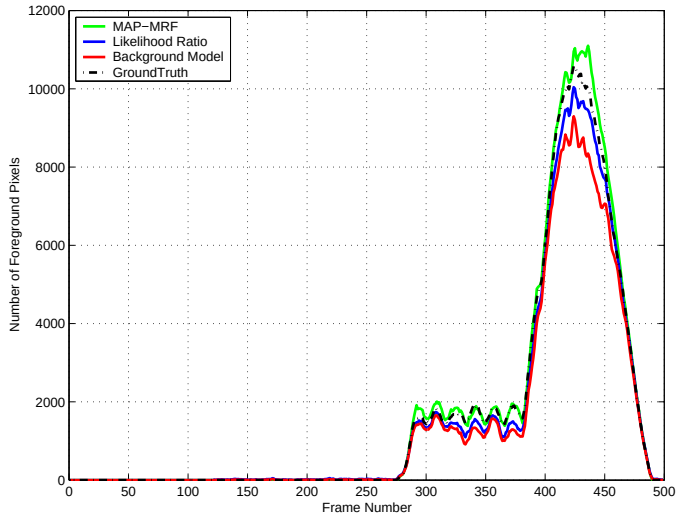
If a contiguous region of pixels was consistently detected corresponding to an object during its period within the field of view, a correct 'object' detection was recorded. If two separate regions were assigned to an object, if an object was not detected or if a region was spuriously detected, a mis-detection was recorded. Results, shown in Table 1, demonstrate that the proposed approach had an overall average detection rate of 99.708% and an overall mis-detection rate of 0.41%. The mis-detections were primarily caused by break-ups in regions, an example of which can be seen in Figure 10(c).

IV. Conclusion

There are a number of innovations in this work. From an intuitive point of view, using the joint representation of image pixels allows local spatial structure of a sequence to be represented explicitly in the modeling process. The entire background is represented by a *single* distribution and a kernel density estimator is used to find membership probabilities. The joint feature space provides the ability to incorporate the spatial distribution of intensities into the decision process, and such feature spaces have been previously used for

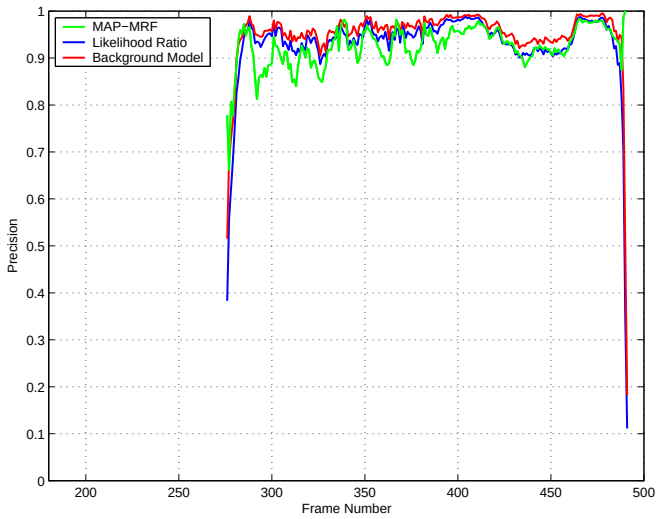


(a)

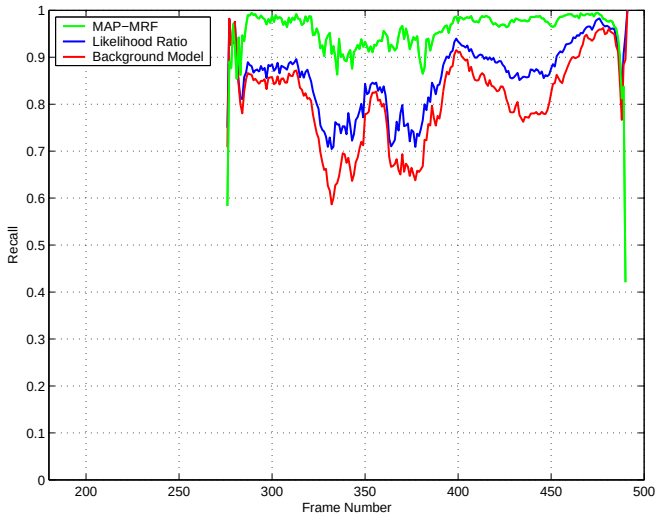


(b)

Fig. 13. Numbers of detected pixels for the sequence with nominal motion (Figure 8). (a) This plot shows the number of pixels detected across each of 500 frames by the Mixture of Gaussians method at various learning rates. Because of the approximate periodicity of the nominal motion, the number of pixels detected by the Mixture of Gaussians method shows periodicity. (b) This plot shows the number of pixels detected at each stage of the proposed approach, (1) using the background model, (2) using the likelihood ratio and (3) using the MAP-MRF estimate.



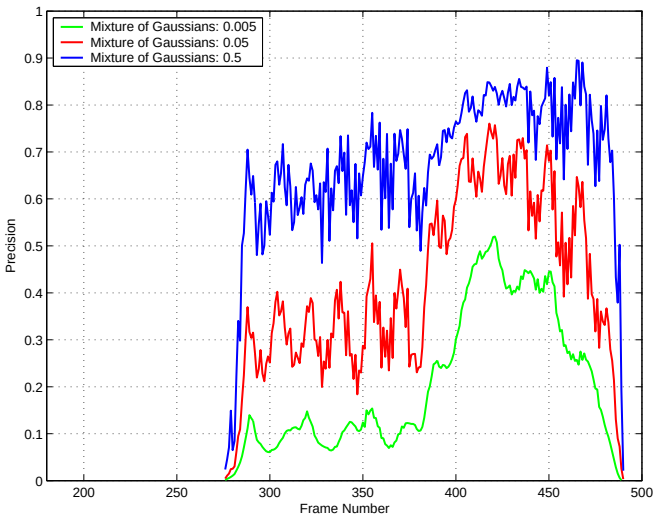
(a)



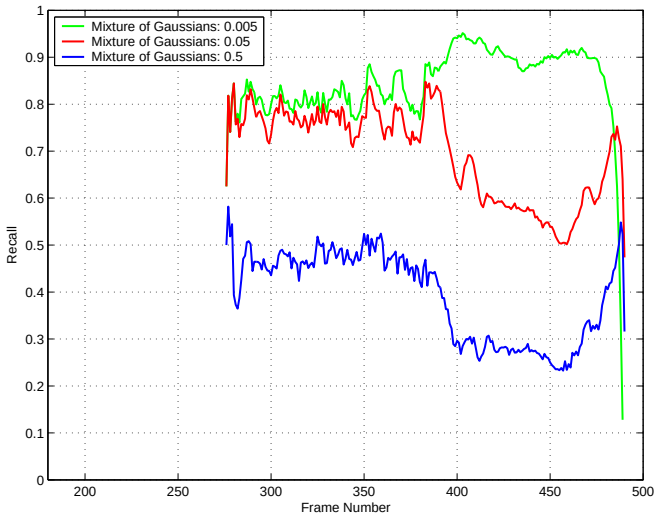
(b)

Fig. 14. Pixel-level detection recall and precision at each level of the proposed approach. (a) Precision and (b) Recall.

image segmentation, smoothing [4] and tracking [6]. A second novel proposition in this work is temporal persistence as a criterion for detection without feedback from higher-level modules (as in [15]). The idea of using both background and foreground color models to compete for ownership of a pixel using the log likelihood ratio has been used before for improving tracking in [3]. However, in the context of object detection, making coherent



(a)



(b)

Fig. 15. Pixel-level detection recall and precision using the Mixture of Gaussians approach at three different learning parameters: 0.005, 0.05 and 0.5. (a) Precision and (b) Recall.

models of both the background and the foreground, changes the paradigm of object detection from identifying outliers with respect to a background model to explicitly classifying between the foreground and background models. The likelihoods obtain are utilized in a MAP-MRF framework that allows an optimal global inference of the solution based on local information. The resulting algorithm performed suitably in several challenging

settings.

Acknowledgments

The authors would like to thank Omar Javed and the anonymous reviewers for their useful comments and advice. This material is based upon work funded in part by the US Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the US Government.

References

[1] J. Besag. On the statistical analysis of dirty pictures. In *Journal of the Royal Statistical Society*, volume 48 of *B*, 1986.

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2001.

[3] R. Collins and Y. Liu. On-line selection of discriminative tracking features. In *IEEE International Conference on Computer Vision*, 2003.

[4] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[5] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

[6] A. Elgammal, R. Duraiswami, and L. Davis. Probabilistic tracking in joint feature-spatial spaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[7] A. Elgammal, D. Harwood, and L. Davis. Background and foreground modeling using non-parametric kernel density estimation for visual surveillance. In *Proceedings of the IEEE*, 2002.

[8] L. Ford and D. Fulkerson. Flows in networks. In *Princeton University Press*, 1962.

[9] N. Friedman and S. Russell. Image segmentation in video sequences: A probabilistic approach. In *Proceedings of the Thirteenth Conference on Uncertainity in Artificial Intelligence*, 1997.

[10] K. Fukunaga. Introduction to statistical pattern recognition. In *Academic Press*, 1990.

[11] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984.

[12] D. Greig, B. Porteous, and A. Seheult. Exact maximum a posteriori estimation for binary images. In *Journal of the Royal Statistical Society*, volume 51 of *B*, 1989.

[13] P. Hall and M. Wand. On the accuracy of binned kernel estimators. In *Journal of Multivariate Analysis*, 1995.

[14] I. Haritaoglu, D. Harwood, and L. Davis. W4: Real-time of people and their activities. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[15] M. Harville. A framework of high-level feedback to adaptive, per-pixel, mixture of gaussian background models. In *Proceedings of the European Conference on Computer Vision*, 2002.

[16] Michael Isard and Andrew Blake. Condensation – conditional density propagation for visual tracking. *Int. J. Computer Vision*, 29(1):5–28, 1998.

[17] R. Jain and H. Nagel. On the analysis of accumulative difference pictures from image sequences of real world scenes. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979.

[18] O. Javed, K. Shafique, and M. Shah. A hierarchical appraoch to robust background subtraction using color and gradient information. In *IEEE Workshop on Motion and Video Computing*, 2002.

[19] M. Jones. Variable kernel density estimates. In *Austrailian Journal of Statistics*, 1990.

[20] K.-P. Karmann, A. Brandt, and R. Gerl. Using adaptive tracking to classify and moitor activities in a site. In *Time Varying Image Processing and Moving Object Recognition*. Elsevier Science Publishers, 1990.

[21] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russell. Towards robust automatic traffic scene analysis in real-time. In *International Conference of Pattern Recognition*, 1994.

[22] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? In *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 2004.

[23] S. Li. Markov random field modeling in computer vision. In *Springer-Verlag*, 1995.

[24] A. Mittal and N. Paragios. Motion-based background subtraction using adaptive kernel density estimation. In *IEEE Proceedings on Computer Vision and Pattern Recognition*, 2004.

[25] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh. Background modeling and subtraction of dynamic scenes. In *IEEE Proceedings of the International Conference on Computer Vision*, 2003.

[26] N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[27] E. Parzen. On estimation of a probability density and mode. In *Annals of Mathematical Statistics*, 1962.

[28] R. Pless, J. Larson, S. Siebers, and B. Westover. Evaluation of local models of dynamic backgrounds. In *IEEE Proceedings on Computer Vision and Pattern Recognition*, 2003.

[29] Y. Ren, C-S. Chua, and Y-K. Ho. Motion detection with nonstationary background. In *Machine Vision and Application*. Springer-Verlag, 2003.

[30] J. Rittscher, J. Kato, S. Joga, and A Blake. A probabilistic background model for trakcing. In *Proceedings of the European Conference on Computer Vision*, 2000.

[31] M. Rosenblatt. Remarks on some nonparametric estimates of a density functions. In *Annals of Mathematical Statistics*, 1956.

[32] S. Sain. Multivariate locally adaptive density estimates. In *Computational Statistics and Data Analysis*, 2002.

[33] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[34] B. Stenger, V. Ramesh, N. Paragios, F Coetzee, and J. Buhmann. Topology free hidden markov models: Application to background modeling. In *Proceedings of the European Conference on Computer Vision*, 2000.

[35] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *IEEE Proceedings of the International Conference on Computer Vision*, 1999.

[36] B. Turlach. Bandwidth selection in kernel density estimation: A review. In *Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin*, 1993.

[37] T. Wada and T. Matsuyama. Appearance sphere: Background model for pan-tilt-zoom camera. *Proceedings of the International Conference on Pattern Recognition*, 1996.

[38] M. Wand and M. Jones. Kernel smoothing. In *Monographs on Statistics and Applied Probability*. Chapman & Hill, 1995.

[39] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland. Pfinder: Real time tracking of the human body. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997.

[40] J. Zhong and S. Sclaroff. Segmenting foreground objects from a dynamic textured background via a robust kalman filter. In *IEEE Proceedings of the International Conference on Computer Vision*, 2003.

[ ]Yaser Sheikh received the B.S. degree in electronic engineering from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan in 2001. He was awarded the Hillman Fellowship in 2004 for excellence in research. Currently, he is working towards his Ph.D. degree at the Computer Vision Laboratory at the University of Central Florida. His current research interests include video analysis, Bayesian inference, human action recognition and co-operative sensing.

[ ]Mubarak Shah is a professor of Computer Science, and the founding director of the Computer Vision Laboratory at University of Central Florida (UCF), is a researcher in computer vision. He is a co-author of two books Video Registration (2003) and Motion-Based Recognition (1997), both by Kluwer Academic Publishers. He has supervised several Ph.D., M.S., and B.S. students to completion, and is currently directing twenty Ph.D. and several B.S. students. He has published close to one hundred fifty papers in leading journals and conferences on topics including activity and gesture recognition, violence detection, event ontology, object tracking (fixed camera, moving camera, multiple overlapping and non-overlapping cameras), video segmentation, story and scene segmentation, view morphing, ATR, wide-baseline matching, and video registration. Dr. Shah is a fellow of IEEE, was an IEEE Distinguished Visitor speaker for 1997-2000, and is often invited to present seminars, tutorials and invited talks all over the world. He received the Harris Corporation Engineering Achievement Award in 1999, the TOKTEN awards from UNDP in 1995, 1997, and 2000; Teaching Incentive Program award in 1995 and 2003, Research Incentive Award in 2003, and IEEE Outstanding Engineering Educator Award in 1997. He is an editor of international book series on "Video Computing"; editor in chief of Machine Vision and Applications journal, and an associate editor Pattern Recognition journal. He was an associate editor of the IEEE Transactions on PAMI, and a guest editor of the special issue of IJCV on Video Computing.