Bayesian Object Detection in Dynamic Scenes

Yaser Sheikh Mubarak Shah

Computer Vision Laboratory
School of Computer Science
University of Central Florida
Orlando, FL 32826

Abstract

Detecting moving objects using stationary cameras is an important precursor to many activity recognition, object recognition and tracking algorithms. In this paper, three innovations are presented over existing approaches. Firstly, the model of the intensities of image pixels as independently distributed random variables is challenged and it is asserted that useful correlation exists in the intensities of spatially proximal pixels. This correlation is exploited to sustain high levels of detection accuracy in the presence of nominal camera motion and dynamic textures. By using a nonparametric 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. Secondly, temporal persistence is proposed as a detection criteria. Unlike previous approaches to object detection which detect objects by building adaptive models of the only background, the foreground is also modeled to augment the detection of objects (without explicit tracking) since objects detected in a preceding frame contain substantial evidence for detection in a current frame. Third, the background and foreground models are used competitively in a MAP-MRF decision framework, stressing spatial context as a condition of pixel-wise labeling and the posterior function is maximized efficiently using graph cuts. Experimental validation of the proposed method is presented on a diverse set of dynamic scenes.

1 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 modelling techniques for the detection of moving objects. 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 and often is 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. 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.

In the context of this work, 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 modelling approaches tend to fall into the first category of pixel-wise models. In their work, Wren *et al* [21] 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 labeled. Similar approaches that used Kalman Filtering for updating were proposed in [8] and [9] and a robust detection algorithm was also proposed in [7]. However, the single Gaussian *pdf* is 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, [2, 18] 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. 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 dynamic 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. 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. Some methods that address the uncertainty of spatial location using local models have also been proposed. In [1], El Gammal *et al* proposed nonparametric estimation methods for per-pixel background modeling. Kernel density estimation (KDE) was used to establish

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

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 dynamic scene by introducing the concept of a spatial distribution of Gaussians (SDG), [16]. 'Nonstationary' backgrounds have most recently been addressed by Pless *et al* [15] and Mittal *et al* [12]. 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 pixel-wise joint-model of color (for a normalized color space), and the optical flow at each pixel. Other notable pixel-wise detection schemes include [19], where topology free HMMs are described and several state splitting criteria are compared in context of background modeling, and [17], where a three-state HMM is used to model the background.

The second category of methods use region models of the background. In [20], 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* [13] used eigenspace decomposition to detect objects.The background was modeled by the eigenvectors corresponding to the η largest eigenvalues, that encompass possible illuminations in the field of view (FOV). 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* [11], Zhong *et al* [22]. 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 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 itself is modeled as a single distribution and multimodal spatial uncertainties are directly handled. Secondly, unlike all 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, without explicitly tracking objects. 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. The rest of the paper is organized as follows. A description of the proposed approach is presented in Section 2. Within this section, a discussion on modelling spatial uncertainty and on utilizing the foreground model for object detection and a description of the overall MAP-MRF framework is included. Experimental results are discussed in Section 3, followed by conclusions in Section 4.

2 Object Detection

In this section we describe the global representation of the background, the use of temporal persistence to formulate object detection as a competitive 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 $\mathcal{S}$.

2.1 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 nonstationary backgrounds in general, the 'correct' model of spatial uncertainty would often have an arbitrary shape and may be bi-modal or multi-modal because by definition, motion follows a certain repetitive pattern. Such arbitrarily structured spaces 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*. Using this representation allows a *global* model of the entire background, $f_{R,G,B,X,Y}(r, g, b, x, y)$, rather than a collection of pixel-wise models. These 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 $\psi_b = \{\mathbf{y}_1, \mathbf{y}_2 \ldots \mathbf{y}_n\}$ 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

([14]). The kernel density estimator is a member of the non-parametric class of estimators 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,

$$P(\mathbf{x}|\psi_b) = n^{-1} \sum_{i=1}^{n} \varphi_{\mathbf{H}}\Big(\mathbf{x} - \mathbf{y}_i\Big), \qquad (1)$$

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

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

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

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

Within the joint domain-range representation, the kernel density estimator explicitly models spatial dependencies, without running into the difficulties of parametric modelling. Furthermore, since it is known that the rgb axes are correlated, it is worth noting that the kernel density estimation also accounts for this correlation. 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.

2.2 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 have smooth motion 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 aspect of this representation is that the foreground model can be constructed in a similar fashion to 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 for the number of foreground samples is defined.

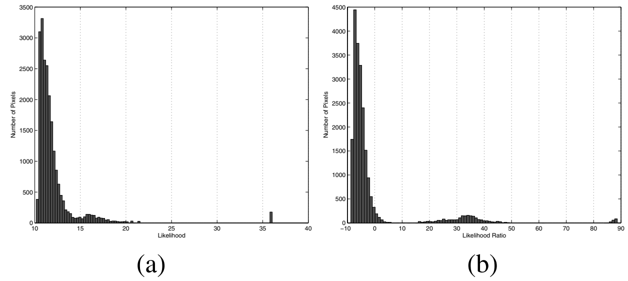


Figure 2: Improvement in discrimination using temporal persistence. (a) Histogrammed log-likelihood values for background membership. (b) Histogrammed log-likelihood ratio values. Clearly the variance *between* clusters is decidedly enhanced.

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

$$P(\mathbf{x}|\psi_f) = \alpha\gamma + (1-\alpha)m^{-1} \sum_{i=1}^{m} \varphi_{\mathbf{H}}\Big(\mathbf{x} - \mathbf{z}_i\Big), \qquad (4)$$

where $\alpha \ll 1$ is a small positive constant that represents the uniform likelihood and γ is the uniform distribution equal to $\frac{1}{R \times G \times B \times M \times N}$ (R, G, B are the support of color values, typically 256, and M, N are the spatial support of the image). If an object is detected in the preceding frame, the likelihood of observing the colors of that object in the same proximity increases according to the second term in Equation 4. Therefore, as objects of interest are detected 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. Figure 1 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*, [4]), $-\ln\frac{P(\mathbf{x}|\psi_b)}{P(\mathbf{x}|\psi_f)} > \kappa$, 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 2. Evidently, the higher the likelihood of belonging to the foreground, the lower the 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.

2.3 MAP-MRF Estimation

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, or by neglecting connected components containing only a few pixels, [18]. Furthermore, directly applying a threshold to membership probabilities implies conditional independence of labels, i.e. $P(\ell_i|\ell_j) = P(\ell_i)$, where $i \neq j$. We assert that

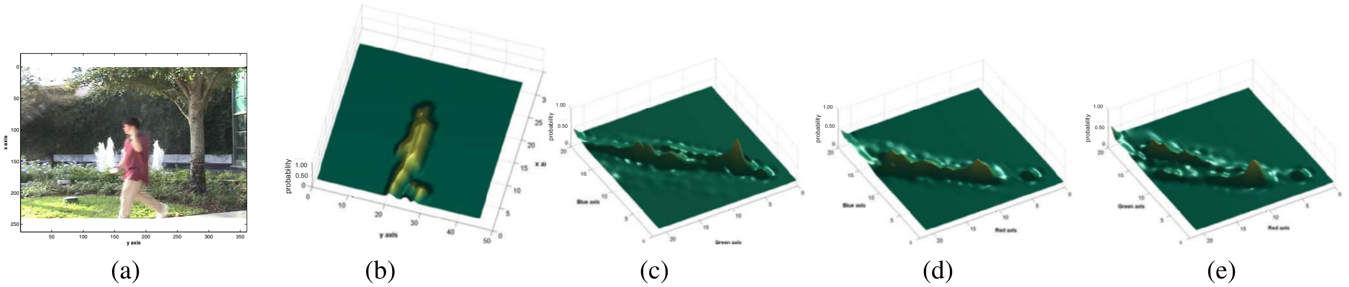


Figure 1: Foreground Modelling. Using kernel density estimates on a model built from recent frames, the foreground can be detected in subsequent frames using the property of temporal persistence, (a) Current Frame (b) the X, Y-marginal, $f_{X,Y}(x, y)$ High membership probabilities are seen in regions 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, $f_{G,R}(g, r)$.

such conditional independence rarely exists between proximal sites. Instead of applying ad-hoc heuristics, Markov Random Fields provide a mathematical foundation to make a global inference using local information. The MRF prior is precisely the constraint of spatial context we wish to impose on $\mathcal{L}$. The set of neighbors, $\mathcal{N}$, is defined as the set of sites within a radius $r \in \mathbb{R}$ from site $\mathbf{i} = (i, j)$,

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

where $distance(\mathbf{a}, \mathbf{b})$ denotes the Euclidean distance between the pixel locations $\mathbf{a}$ and $\mathbf{b}$. The 4-neighborhood or 8-neighborhood cliques are two commonly used neighborhoods. The pixel-vectors $\hat{\mathbf{x}} = \{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_p\}$ are conditionally independent given $\mathcal{L}$, with conditional density functions $f(\mathbf{x}_i|\ell_i)$. Thus, since each $\mathbf{x}_i$ is dependant on $\mathcal{L}$ only through ℓ_i, the likelihood function may be written as,

$$l(\hat{\mathbf{x}}|\mathcal{L}) = \prod_{i=1}^{p} f(\mathbf{x}_i|\ell_i) = \prod_{i=1}^{p} f(\mathbf{x}_i|\psi_f)^{\ell_i} f(\mathbf{x}_i|\psi_b)^{1-\ell_i} \quad (5)$$

Spatial context is enforced in the decision through a pairwise interaction MRF prior, used for its discontinuity preserving properties, $p(\mathcal{L}) \propto \exp\big(\sum_{i=1}^{p}\sum_{j=1}^{p} \lambda\big(\ell_i\ell_j + (1-\ell_i)(1-\ell_j)\big)\big)$, where λ is a constant, and $i \neq j$ are neighbors. By Bayes Law,

$$p(\mathcal{L}|\hat{\mathbf{x}}) = \frac{p(\hat{\mathbf{x}}|\mathcal{L})p(\mathcal{L})}{p(\hat{\mathbf{x}})} \quad (6)$$

where $p(\hat{\mathbf{x}}|\mathcal{L})$ is as defined in Equation 5, $p(\mathcal{L})$ is as defined and $p(\hat{\mathbf{x}}) = p(\hat{\mathbf{x}}|\psi_f) + p(\hat{\mathbf{x}}|\psi_b)$. The log-posterior, $\ln p(\mathcal{L}|\hat{\mathbf{x}})$, is then equivalent to (ignoring constant terms),

$$L(\mathcal{L}|\hat{\mathbf{x}}) = \sum_{i=1}^{p} \ln\left(\frac{f(\mathbf{x}_i|\psi_f)}{f(\mathbf{x}_i|\psi_b)}\right)\ell_i + \sum_{i=1}^{p}\sum_{j=1}^{p} \lambda\Big(\ell_i\ell_j + (1-\ell_i)(1-\ell_j)\Big). \quad (7)$$

The MAP estimate is the binary image that maximizes

$$\arg\max_{\mathcal{L}\in\mathfrak{L}} L(\mathcal{L}|\hat{\mathbf{x}}) \quad (8)$$

	Objects	Det.	Mis-Det.	Det. Rate	Mis-Det. Rate
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 1: Object level detection rates. Object sensitivity and specificity for five sequences (each one hour long).

where $\mathfrak{L}$ are the 2^{NM} possible configurations of $\mathcal{L}$. An exhaustive search of the solution space is not feasible due to its size, but since L belongs to the $\mathcal{F}^2$ class of energy functions (as defined in [10]), efficient algorithms exist for the maximization of L using graph cuts, [5, 10]. To optimize the energy function (Equation 7), we construct a graph $\mathcal{G} = \langle\mathcal{V}, \mathcal{E}\rangle$ with a 4-neighborhood system $\mathcal{N}$. 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 $\mathcal{V} = \{v_1, v_2, \cdots, v_n, s, t\}$. The graph construction is as described in [5], with a directed edge (s, i) from s to node i with a weight τ (the log-likelihood ratio), if $\tau > 0$, otherwise a directed edge (i, t) is added between node i and the sink t with a weight τ. For the second term in Equation 7, undirected edges of weight λ are added if there corresponding pixels are neighbors as defined by $\mathcal{N}$. The minimum cut can then computed through several approaches, the Ford-Fulkerson algorithm [3], the faster version in [5] or through the generic version of [10]. The configuration found corresponds to an optimal estimate of $\mathcal{L}$.

3 Results and Discussion

The algorithm was tested 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 can process up to 11 fps for a frame size of 240 by 360. Comparative results for the mixture of Gaussians method have also been shown. 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 in the scene. In Figure 4 the first row is the current image. The second row shows the detected fore-

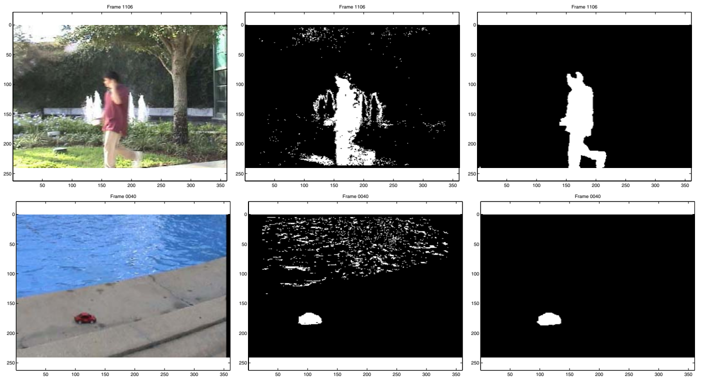


Figure 3: Detection in dynamic scenes. The first column has the original images, the second column shows the results obtained by the Mixture of Gaussians method, [18] and the third column are the results obtained by the proposed method. Morphological operators were not used in the results.

ground proposed in [18], and it is evident that the motion causes substantial degradation in performance, despite a 5-component mixture model and a 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. Figures 3 shows results on a variety of scenes with dynamic textures, including fountains (a), shimmering water (b) and waving trees (c) and (d).

We performed quantitative analysis at both the pixel-level and object-level. For the first experiment, we manually segmented a 300-frame sequence containing nominal motion (as seen in Figure 4). In the sequence, two objects (a person and then a car) move across the field of view causing the two bumps in the number of pixels. The per-frame detection rates are shown in Figure 5 in terms of specificity and sensitivity, where

$$\text{specificity} = \frac{\#\text{ of true positives detected}}{\text{total }\#\text{ of true positives}}$$

$$\text{sensitivity} = \frac{\#\text{ of true negatives detected}}{\text{total }\#\text{ of true negatives}}.$$

Clearly, the detection accuracy both in terms of sensitivity and specificity is consistently higher than the mixture of Gaussians approach. Next, to evaluate detection at the object level (detecting whether an object is present or not), we evaluated five sequences, each one hour long. Sensitivity and specificity were measured in an identical fashion to the pixel-level experiment, with an object as each contiguous region of pixels. Results are shown in Table 1.

4 Conclusion

There are a number of fundamental 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.

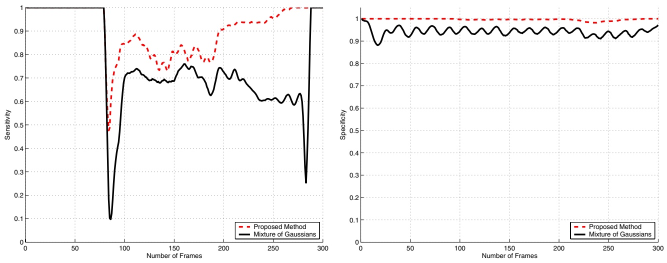


Figure 5: Pixel-level detection sensitivity and specificity. Average True Negatives - Proposed Method 99.65 %, Average True Negatives - Mixture of Gaussians 94.22 %, Average True Positives - Proposed Method 90.66 %, Average True Positives - Mixture of Gaussians 75.42 %

The background is represented by a *single* distribution and a kernel density estimator is to find membership probabilities. Another novel proposition in this work is temporal persistence as a criterion for detection without feedback from higher-level modules. By making coherent 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 in this way 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.

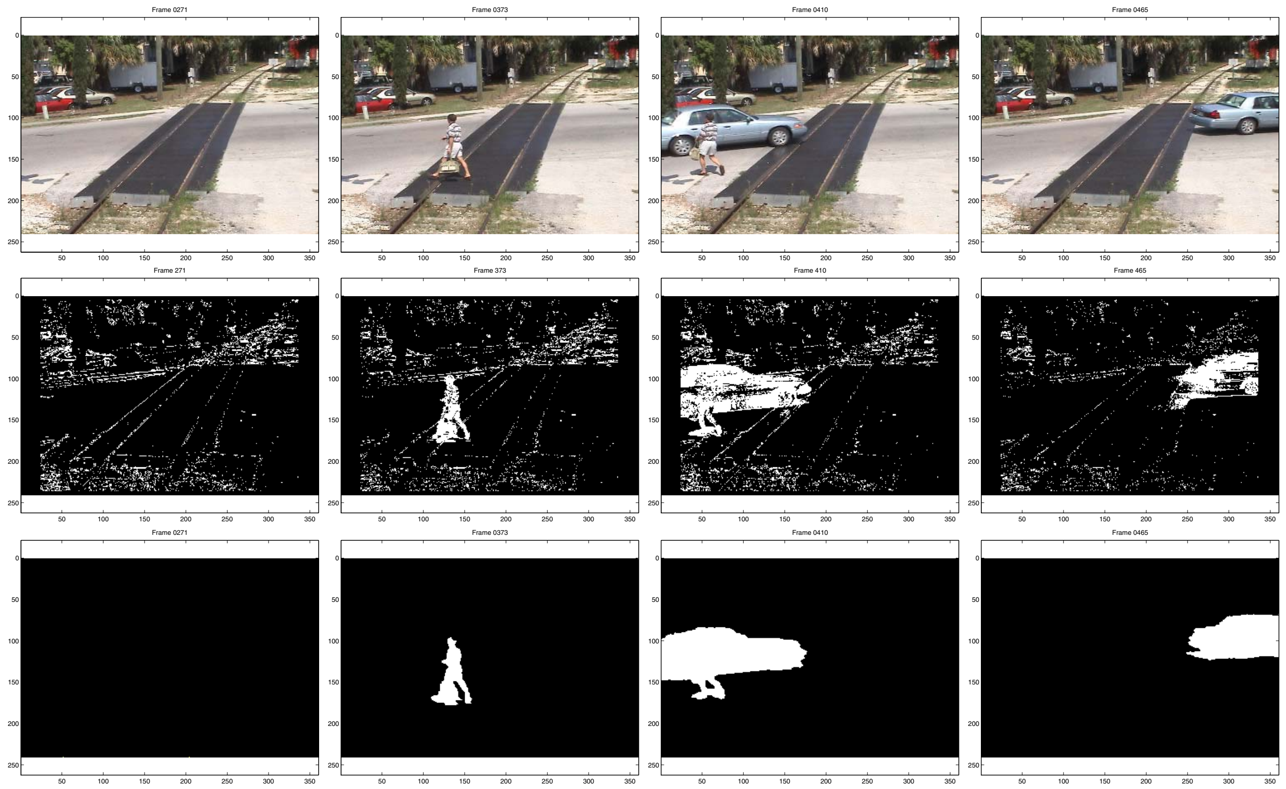
Since analysis is being performed in $\mathbb{R}^5$, it is important to consider how the so-called curse of dimensionality affects performance. Typically higher dimensional feature spaces mean large sparsely populated volumes, but at high frame rates, the overriding advantage in the context of background modeling and object detection is the generous availability of data. Here, the magnitude of the sample size is seen as an effective means of reducing the variance of the density estimate, otherwise expected [4] (pg. 323). Future directions include using a fully parameterized bandwidth matrix for use in adaptive Kernel Density Estimation. Another promising area of future work is to fit this work in with nonparametric approaches to tracking, like mean-shift tracking. Since both background and foreground models are continuously maintained, the detection information can be used to weight likelihoods *apriori*.

Acknowledgements

The authors would like to thank Omar Javed for his comments and critiques. This material is based upon work funded in part by the U. S. 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 U.S. Government.

References

[1] A. Elgammal , D. Harwood and L. Davis, *"Background and Foreground Modeling Using Nonparametric Kernel Density Estimation for Visual Surveillance,"* IEEE Proceedings, 2002.



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

Figure 4: 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 the Mixture of Gaussians method, [18] and the third row results obtained by the proposed method. Morphological operators were not used in the results.

[2] N. Friedman and S. Russell, *"Image Segmentation in Video Sequences: A Probabilistic Approach,"* Proceedings of the Thirteenth Conference on Uncertainity in Artificial Intelligence, 1997.

[3] L. Ford, D. Fulkerson, *"Flows in Networks"*, Princeton University Press, 1962.

[4] K. Fukunaga, *"Introduction to Statistical Pattern Recognition"*, Academic Press, 1990.

[5] D. Greig, B. Porteous, A. Seheult, *"Exact Maximum A Posteriori Estimation for Binary Images,"* Journal of the Royal Statistical Society, Series B, Vol. 51, No. 2, 1989.

[6] S. Geman and D. Geman, *"Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images,"* TPAMI, 1984.

[7] I. Haritaoglu, D. Harwood and L. Davis, *"W4: Real-time of people and their activities"*, TPAMI, 2000.

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

[9] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russell, *"Towards robust automatic traffic scene analysis in real-time,"* ICPR, 1994.

[10] V. Kolmogorov and R. Zabihm *"What Energy Functions can be Minimized via Graph Cuts?,"* TPAMI, 2004.

[11] A. Monnet, A. Mittal, Nikos Paragios, and V. Ramesh, *"Background Modeling and Subtraction of Dynamic Scenes,"* ICCV, 2003.

[12] A. Mittal and N. Paragios, *"Motion-Based Background Subtraction using Adaptive Kernel Density Estimation,"* CVPR, 2004.

[13] N. Oliver, B. Rosario, and A. Pentland, *"A Bayesian Computer Vision System for Modeling Human Interactions,"* TPAMI, 2000.

[14] E. Parzen, *"On Estimation of a Probability Density and Mode,"* Annals of Mathematical Statistics, 1962.

[15] R. Pless, J. Larson, S. Siebers and B. Westover, *"Evaluation of Local models of Dynamic Backgrounds,"* CVPR, 2003.

[16] Y. Ren, C-S. Chua and Y-K. Ho, *"Motion Detection with Nonstationary Background,"* MVA, Springer-Verlag, 2003.

[17] J. Rittscher, J. Kato, S. Joga, and A Blake. *"A probabilistic background model for tracking,"* ECCV, 2000.

[18] C. Stauffer and W. Grimson, *"Learning Patterns of Activity using Real-time Tracking,"* TPAMI, 2000.

[19] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee and J. Buhmann. *"Topology Free Hidden Markov Models: Application to Background Modeling,"* ECCV, 2000.

[20] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, *"Wallflower: Principles and Practice of Background Maintenance,"* ICCV, 1999.

[21] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland, *"Pfinder: Real time Tracking of the Human Body,"* TPAMI, 1997.

[22] J. Zhong and S. Sclaroff, *"Segmenting Foreground Objects from a Dynamic Textured Background via a Robust Kalman Filter,"* ICCV, 2003.