# DYNAMIC SCENE MODELING FOR OBJECT DETECTION USING SINGLE-CLASS SVM

*Imran N. Junejo*

University of Sharjah, Sharjah, U.A.E.

*Adeel A. Bhutta* *Hassan Foroosh*

University of Central Florida
Orlando, FL 32816, U.S.A.

## ABSTRACT

Scene modeling is the starting point and thus the most crucial stage for many vision based systems involving tracking, or recognition. Most of the existing approaches attempt at solving this problem by making some simplifying assumptions such as that of a stationary background. However, this is not always the case, as swaying trees, or ripples in the water etc often violate these assumptions. In this paper, we present a novel method for modeling a *dynamic* scene, i.e. scenes that contain "non-stationary" background motions, such as periodic motions (e.g. pendulums or escalators) or dynamic textures (e.g. water fountain in the background, swaying trees, or water ripples etc). The proposed method introduces single-class Support Vector Machines (SVM), and we show why it is preferable to other scene modeling techniques currently in use for this particular problem. Using a rectangular region around a pixel, spatial and appearance based features are extracted from the images, used for learning the SVMs. We experiment on a diverse set of dynamic scenes and present both qualitative and quantitative results; indicating the practicality and the effectiveness of the proposed method.

***Index Terms***— background subtraction, scene modeling, dynamic scene, single-class classification

## 1. INTRODUCTION

Intelligent video surveillance has attracted immense attention from both the researchers and the industry at large. Applications such as human tracking, surveillance, segmentation and automatic analysis of foreground objects are some of the key applications. Most of these tasks, however, involve background modeling. The goal is to extract objects of interest, generally assumed to be *non-stationary* with respect to the created scene model. Efficient and a robust solution to this initial phase holds the key to an improved vision based system performing meaningful higher level tasks.

Surveillance systems typically use stationary cameras to monitor an area of interest. Stationarity of the sensor has been the key assumption that has allowed researcher to employ various statistical techniques for scene modeling. It is assumed that generally an *interesting* object will be moving or non-stationary compared to the static scene [8]. An accurate tracking relies on a reliable detection of such objects [3, 11]. However, this assumption is often violated in real-world scenarios where many factors, such as windy conditions, often cause the sensor to move or sway slightly. In addition, a stationary sensor does not guarantee a stationary *background*. Examples of such backgrounds includes water ripples, swaying trees, moving fan, or a water fountain in the background. It is these periodic or recurrent motions that cause non-stationarity in the scene and prompts the existing methods to recognize these phenomena as interesting.

In order to model a dynamic scene, the research gradually moved towards modeling each pixel in the scene by a single Gaussian distribution. [8] in their seminal work, fitted a single three dimensional Gaussian per pixel, where the model parameters, i.e. mean and the standard deviation, was estimated from the pixels in consecutive frames. However, one Gaussian model per pixel has proved to be ill-suited to capture the different underlying processes generating the pixel intensity in an outdoor scene. [7] thus proposed modeling the scene with a mixture of Gaussians (MoG). An incident pixel was classified to be background or foreground based on its similarity to every Gaussian density of that pixel's model learned from the scene. If a match is found, the mean and the variance of the matched Gaussian is updated, otherwise a new Gaussian is created with some predefined initial parameters. A non-parametric kernel density estimation (KDE) was adopted by [2] for per pixel background modeling. Both of these methods address nominal camera moments but to a large extent have been unable to handle larger phenomena generating from the background. See [1] for a review of other methods.

However, the work most relevant to that of ours is that of: [5, 6, 10]. In their work, [10] construct a covariance matrix from the region surrounding a pixel by using the spatial and the appearance attributes. The covariance matrix from a new pixel is matched to that of the constructed model to distinguish a foreground from the background using certain threshold values. [5] addresses the non-stationary background and used on-line auto-regressive models for scene modeling. [6] models the scene using a single probability density in addition to modeling the foreground. They introduce a *temporal persistence* for accurate detection. The maintained background and the foreground models are used in a MAP-MRF selection framework for performing the object detection in a stationary

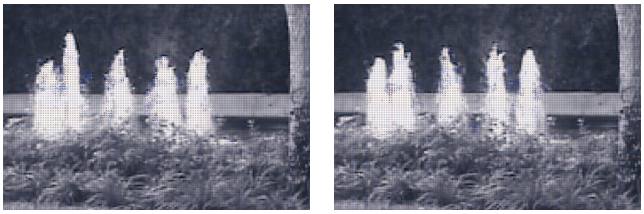

**Fig. 1**. Sample images from the `fountain` sequence showing the flow vectors super-imposed on the original image. Sequences such as these, contain dynamic and periodic motions, creating in inaccuracies for the traditional approaches.

camera, where the maximum a posteriori solution is obtained by finding the minimum cut of the constructed graph.

In this paper, we treat the scene or the background modeling problem as a Single-Class Classification (SCC) problem and introduce the single-class SVM for scene modeling [9]. SCC aims to distinguish one class of data from the universal set of multiple classes. Without requiring a large amount of data, Single-Class SVM classify one class of data from the rest of feature space given only positive data by drawing a optimum non-linear boundary of the positive data set in the feature space. In addition we use a novel set of region based features to capture the dynamics of the background. These proposed features not only capture the dynamics at each pixel, they also capture the spatial context of the region surrounding a pixel.

The rest of the paper is organized as follows: The scene modeling step, involving feature extraction, and single-class classification using SVM is defined in Section 2. The experiments performed on a sequence and the obtained results are presented in Section 3, followed by conclusion.

## 2. SCENE MODELING

In a dynamic scene, every pixel in the image is undergoing a certain periodic or a repetitive change in intensities at each time instance. It is too simplistic to assume that a pixel intensity varies independently of its neighbors [6]. For example, in a typical scene with swaying trees or water ripples, a larger region of the image, not just a single pixel, is involved in the same type of motion. At the same time, there is a temporal continuity in the motion, as in the case of swaying trees, where branches or the leaves move back and forth. Thus it is essential that both the spatial and the temporal context be captured for an accurate scene modeling.

For a given set of images $\{\mathbf{I}(t)\}_{t=1\ldots k}$, we first compute the optical flow by using Lucas and Kanade method [4] on the whole image using two consecutive frames and generate four representation, i.e. the $v_x$ and $v_y$ components of optical flow, the angle $\theta$ between each corresponding flow components and the magnitude $\mathfrak{m}$ for the $v_x$ and $v_y$ components such that: $\mathcal{F} = \{v_x, v_y, \mathfrak{m}, \theta\}$. The idea is to extract a set of features that uniquely capture the dynamics of the scene by using these four representations.

### 2.1. Feature Set

Once we have computed the optical flow, for every pixel in the image, denoted by $p_t^i$ i.e. the $i^{th}$ pixel in image $t$, a rectangular region of the size $S$ is used to compute the following set of simple features:

**Entropy:** The standard way of defining entropy is,

$$h = \sum^{M}\sum^{N} \mathcal{F}_i \log(\mathcal{F}_i) \tag{1}$$

where $\mathcal{F}_i \in \{v_x, v_y, \mathfrak{m}, \theta\}$ for the $i^{th}$ pixel, and $M$ and the $N$ is the $M \times N$ region around the pixel. Generally this is set to be $5 \times 5$ in our experiments. The entropy $h$ is a statistical measure of the randomness that can be used to characterize the flow vectors, magnitude or the angle between the flow vectors i.e. $\mathcal{F}_i$.

**Energy:** The amount of energy in an $M \times N$ region surrounding a pixel is computed as:

$$e = \sum^{M}\sum^{N} (\mathcal{F}_i)^2 \tag{2}$$

where $\mathcal{F}_i$ is as defined above. $e$ measures the energy present in the flow vectors, magnitude or the angle between the flow vectors in an $M \times N$ region around a pixel.

**Inertia:** Finally, we define the inertia as,

$$j = \sum_{u=1}^{M}\sum_{v=1}^{N} (u-v)^2 \mathcal{F}_i \tag{3}$$

where $j$ measures an object's resistance to changes in its rotation rate.

The features defined above are unique, and yet simple to compute. Entropy, inertia and energy are relatively immune to *rotation*, since the order is not important. These measures are *scale* invariant, and are inherently invariant to linear change in *illumination* as well. The output at this stage is a 12-dimensional feature vector
$H^{p_t^i} = \{h_{v_x}, e_{v_x}, j_{v_x}, h_{v_y}, e_{v_y}, j_{v_y}, h_\theta, e_\theta, j_\theta, h_\mathfrak{m}, e_\mathfrak{m}, j_\mathfrak{m}\}$ for every pixel $p_t^i$ in the frame $t$.

### 2.2. Single-Class Classification

The scene modeling problem involves observing a scene which is assumed to contain an acceptable behavior. During this phase, which is generally termed as the training phase, it is possible to only gather the *positive data* that describes what belongs to the scene. However, during this phase it

is not possible to include the negative data which is later to be detected as such. This scenario is a good candidate for applying the Single Class Classification techniques.

Given a limited number of training data, the *optimal* class boundary function is considered the one that gives the best generalization performance which represents the performance on unseen examples. For supervised learning, SVM tries to maximize the generalization by maximizing the margin and also supports nonlinear separation using advanced kernels; thus avoiding underfitting and overfitting [9].

More specifically, we adopt the SVMC as proposed by [9], which employs the Mapping Convergence framework where the algorithm generates the boundary close to the optimum. As the sample size increases, SVMC prevents training time from increasing dramatically, and the running time is shown to be asymptotically equal to that of a SVM. The approach is to use minimally required data at each iteration so that the data does not degrade the accuracy of the boundary. In their work, [9] prove that the training time is $O(n^2)$, where $n$ is the size of the training data. Thus for training on data set of size $K$ images, we compute the feature vector $H^{p^i} = \{H^{p_1^i}, H^{p_2^i}, ..., H^{p_K^i}\}$ for each pixel location. This feature vector is used to train the SVMC at each pixel location.

SVMC has been shown to have a good accuracy for single class classification by computing accurate classification boundary around the positive data (during the training phase) using the unlabeled data in a systematic way. Moreover, SVMC does not require a large amount of positive training data while still maintaining performance close to that of original SVM while providing good generalization, as the results in the next section show.

## 3. EXPERIMENTS AND RESULTS

We tested the proposed method on a publicly available `fountain` sequence. The images have a resolution of $320 \times 240$. The sequence contains nominal camera motion, and significant dynamic textures and cyclic motion. Dynamic texture is induced in the scene by the moving trees while the fountain in the background induces constant cyclic motion. We also compare the proposed method with the Mixture of Gaussian approach [7], trained using three component mixture model and 400 frames for training purpose. For our method, we only used seventy five frame for feature extraction and for training the Single-class SVM, as described in the Section 2.

Qualitatively, the results are an improvement over the method [7], as shown in Figure 2. The camera is mounted on a tall tripod, and the wind causes the tripod to sway back and forth; and in the background is a water fountain and swaying trees. The first row shows the original images from the test sequence. The figure depicts a person comings in from the left of the image and walking to the right. The second row shows the results obtained from the MoG method and it become evident that the nominal motion caused by the camera, and the presence of the water fountain, cause substantial degradation of the results qualitatively. A large number of moving background pixels are detected as foreground pixels. Some portions of the foreground object are also classified as background. The third row shows the results obtained by the proposed method, showing a considerable improvement over MoG [7]. The fourth row shows the ground truth frames obtained by manually labeling some frames from the image sequence.

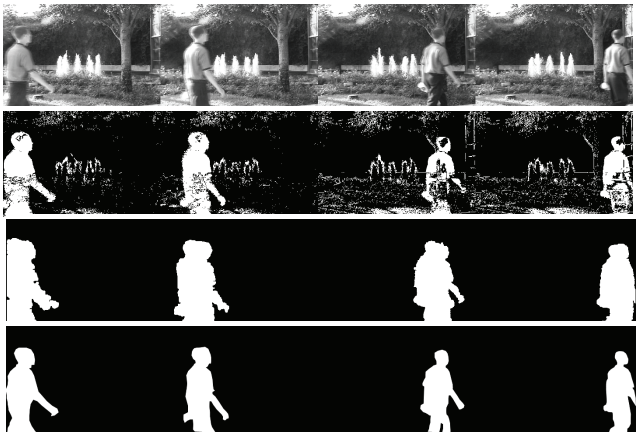

**Fig. 2**. Experimental results obtained from the fountain sequence. The first row shows the original images followed by the results obtained by the Mixture of Gaussian approach, shown in row 2 (trained on 400 images). The results obtained from the proposed method are shown in row 3 and the ground truth subtraction results are shown in the last row.

Quantitative analysis is performed on the sequence and the results obtained from our method are compared to [7], shown in Figure 3. We compute the following two measures for assessing the goodness of the proposed method:

$$\texttt{Precision} = \frac{\texttt{\# of true positives detected}}{\texttt{total \# of positives detected}}$$

$$\texttt{Recall} = \frac{\texttt{\# of true positives detected}}{\texttt{total \# of true positives}}$$

The detection accuracy, in terms of both the precision and the recall is considerably higher than the mixture of Gaussian approach. As observed from the figure, the recall rate for the proposed method is consistently high, whereas at some instances the precision decreases due to strong motions in the image sequences. This indicates that the localized foreground is larger than the labeled ground truth, however, the background pixels such as the fountain and the swaying trees are

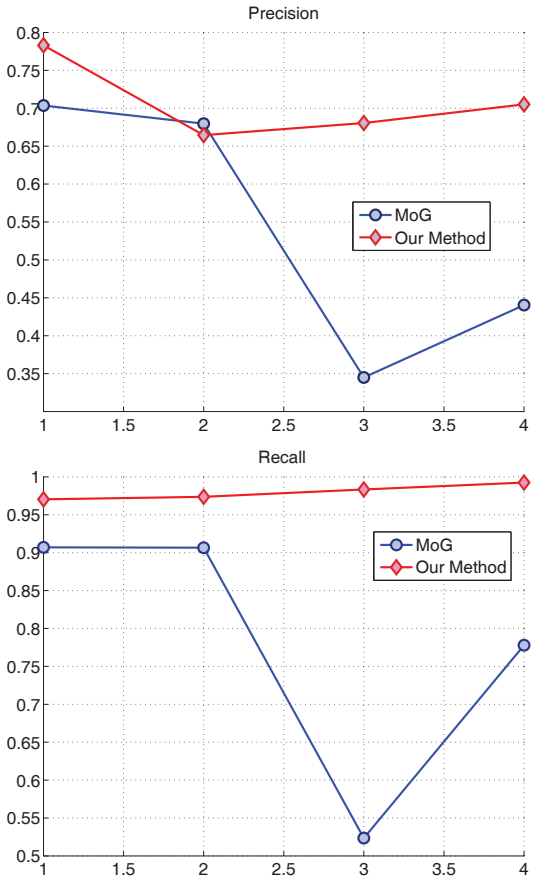


**Fig. 3**. Comparison of the proposed method with the state of the art MoG method. The top figure shows the calculated precision for the two methods while the bottom figure shows the computed recall for the two methods. These figures indicate that the precision and recall of the proposed method is considerably better than the standard MoG approach.

not detected as foreground objects at all. Moreover, we are not performing any post-processing techniques, such as graph cuts [6] to improve the boundaries of the foreground object, which would improve the precision considerably.

## 4. CONCLUSION

Scene modeling is a very significant initial step for various vision based systems. The existing methods often fail for scenes with dynamic textures or cyclic background motion. We propose treating the scene modeling problem as a Single-Class classification problem, and propose using single-class SVM that is able to create the optimal class boundary from a very limited set of training example. We also employ a novel, yet simple region based features, extracted at each pixel location for training the single-class SVMs. The proposed features not only capture the dynamics at each pixel, they also capture the spatial context of the region surrounding a pixel. We have presented results on a challenging sequence containing considerable amount of sensor motion, in addition to a dynamic background. We compare our results with the standard Mixture of Gaussians approach and notice a very significant improvement. The proposed method has successfully minimized false positives and shows considerably higher recall and precision compared to the MoG approach. These encouraging results indicate the practicality of the proposed method.

## 5. REFERENCES

[1] Y. Benezeth, P. Jodoin, B. Emile, H. Laurent, and C. Rosenberger. Review and evaluation of commonly-implemented background subtraction algorithms. In *Proc. International Conference on Pattern Recognition*, pages 1–4, 2008.

[2] A. Elgammal, R. Duraiswami, D. Harwood, L. S. Davis, R. Duraiswami, and D. Harwood. Background and foreground modeling using nonparametric kernel density for visual surveillance. In *Proceedings of the IEEE*, pages 1151–1163, 2002.

[3] O. Javed and M. Shah. Tracking and object classification for automated surveillance. In *the seventh European Conference on Computer Vision (ECCV)*, 2002.

[4] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Image Understanding Workshop*, pages 121–130, 1981.

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

[6] Y. Sheikh and M. Shah. Bayesian modeling of dynamic scenes for object detection. *PAMI*, 27(11):1778–1792, November 2005.

[7] C. Stauffer, W. Eric, and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:747–757, 2000.

[8] C. Wren, A. Azarbayejani, T. Darell, and A. Pentland. Pfinder: Real-time tracking of human body. *PAMI*, 19:780–785, 1997.

[9] H. Yu. Single-class classification with mapping convergence. *Mach. Learn.*, 61(1-3):49–69, 2005.

[10] S. Zhang, H. Yao, S. Liu, X. Chen, and W. Gao. A covariance-based method for dynamic background subtraction. In *ICPR*, pages 1–4, 2008.

[11] T. Zhao, M. Aggarwal, R. Kumar, and H. Sawhney. Real-time wide area multi-camera stereo tracking. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2005.