# Hierarchical Background Subtraction using Local Pixel Clustering

Bineng Zhong[1], Hongxun Yao[1], Shiguang Shan[2,3], Xilin Chen[2,3], Wen Gao[1, 2,4]

1Department of Computer Science and Engineering, Harbin Institute of Technology,
2Digital Media Research Center, Institute of Computing Tech., CAS, Beijing, China
3Key Laboratory of Intelligent Information Processing, CAS, Beijing,China
4Digital Media Institute, Peking University, Beijing, China
{bnzhong, sgshan, xlchen, wgao }@jdl.ac.cn; yhx@vilab.hit.edu.cn

## Abstract

*We propose a robust hierarchical background subtraction technique which takes the spatial relations of neighboring pixels in a local region into account to detect objects in difficult conditions. Our algorithm combines a per-pixel with a per-region background model in a hierarchical manner, which accentuates the advantages of each. This is a natural combination because the two models have complementary strengths. The per-pixel background model is achieved by mixture of Gaussians Models (GMM) with RGB feature. Although precisely describing background change in high resolution, it suffers from the sensitivity to quick variations in dynamic environment. To tolerate these quick variations, we further develop a novel GMM based per-region background model, which is updated by the cluster centers obtained from a k-means clustering of the pixels' RGB feature in the region. Numerical and qualitative experimental results on challenging videos demonstrate the robustness of the proposed method.*

## 1. Introduction

Background subtraction (BGS) is a basic problem in intelligence video surveillance. Challenges in the BGS problem include: (1) Background object motions, for instance, tree leaves swaying and water rippling. (2) Illumination variations due to sunlight and weather changes. (3) Camera jitters due to support vibration, wind, etc. To achieve desirable foreground detection even in such a dynamic environment (see Fig. 3(a)), a large number of BGS methods have been proposed over the years.

One popular technique is to model each pixel feature in a video frame with the mixture of Gaussians Models (GMM) [1]. The GMM method can deal with periodic motions from a cluttered background, slow lighting changes, etc. However, it cannot adapt to the quick variations in dynamic environments [2]. Other pixel-wise modeling techniques include kernel density estimation [3, 4] and code-book [5], etc. Although the above BGS methods have significantly different modeling schemes, most of them share the same basic assumption that the time series of observations is independent on each pixel which is a strict assumption and limit their application in dynamic environment. Thus, some methods jointly using pixel and region models have been proposed. In [2, 6], a 3-stage algorithm is separately presented, which operates respectively at pixel, region and frame level. In [7], a Kalman filter is used for modeling image regions as an autoregressive moving average process. It works well for periodical changes in a scene but it is difficult to predict background changes with varying frequency in the natural scene. After modeling the background with GMM, [8] integrates intensity and texture information to remove shadows and to enable the algorithm working for quick lighting changes. In [9], scene is coarsely represented as the union of pixel layers and foreground objects are detected by propagating these layers using a maximum-likelihood assignment. However, the limitations of the method are high-computational complexity and the requirement of an extra offline training step. Please refer to [10] for a more complete BGS methods review.

In this paper, we propose a robust hierarchical BGS technique which takes advantages of meaningful correlation between pixels in the spatial vicinity to detect objects in difficult conditions. Our algorithm combines a per-pixel background model with a per-region background model in a hierarchical manner that accentuates the advantages of each. This is a natural combination because the two models have complementary strengths. In the pixel processing level, background is updated by the traditional GMM model.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

Although it can precisely adapt to the background gradual change, it is sensitive to quick variations in dynamic environment, such as background object motions, illumination variations, and camera jitter. To make our BGS more robust to these quick variations, we further provide a novel GMM based region level background modeling mechanism. In this region-wise processing, GMM is updated by the cluster centers obtained from a k-means clustering of the pixels' RGB feature in the region. Since our method utilizes the spatial property of background image variations, it is not affected by the quick image variations. Numerical and qualitative experimental results on challenging videos demonstrate the robustness of the proposed method.

The rest of the paper is organized as follows: Section 2 describes the proposed method in details. In section 3, experimental results are given. Finally, conclusion and future work are drawn in Section 4.

## 2. The Proposed Method

### 2.1. The Per-pixel Background Model

We use the popular GMM [1] for per-pixel background modeling. In the GMM, three significant parameters $K_p$, $T_p$ and $\alpha_p$ are needed to be set, where $K_p$ is the number of Gaussian components, $T_p$ the minimum portion of the background model and $\alpha_p$ the learning rate. In our implementation, per-pixel background modeling is done in terms of RGB color and we set $K_p = 3$ (three Gaussians), $T_p = 0.7$ and $\alpha_p = 0.01$.

The GMM method can deal with periodic motions from a cluttered background, slow lighting changes, etc. However, it cannot adapt to the quick variations in dynamic environments, such as tree leaves swaying, water rippling, and camera jitter. In other words, the GMM method generates large number of false foreground pixels under those difficulty conditions (see Fig. 3(a)).

### 2.2. The Per-region Background Model

We use the per-region background model to determine whether the foreground pixel detected by per-pixel processing obeys the learned local region background model. The basic idea is to utilize the spatial correlation of background image variations. To do this, for each pixel, we define a local region of size $N \times N$ centered at it. Here N is a user-preferred parameter that indicates the spatial variance of the current pixel. Each local region is modeled by a GMM and the GMM is adaptively updated with information in the corresponding region from new frames. The algorithm of per-region background model is depicted in Fig.1.

Specifically, we adopt a 3-component GMM for each pixel's per-region background model:

$$P_{region}(X_t) = \sum_{i=1}^{3} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \quad (1)$$

where $\eta$ is a Gaussian probability density function

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X_t - \mu_t)^T \Sigma^{-1}(X_t - \mu_t)\right) \quad (2)$$

and $\omega_{i,t}$, $\mu_{i,t}$ and $\Sigma_{i,t}$ are the time adaptive mixture coefficients, mean and variance, respectively, of the ith Gaussian of the mixture. We first perform a k-means clustering with $K_{cluser} = 3$ on pixels in the local region. Then, the per-region background model $P_{region}$ is adaptively update by performing similar learning schema as [1] on the means of each cluster's color, instead of every pixel in the local region, which can largely reduce the number of model updating. In our experiments, for $P_{region}$, the minimum portion of the background model $T_r$ is set as 0.8 to allow more portion of the background model, the learning rate $\alpha_r$ is adaptively adjusted as described in the following subsection, and the size of neighborhood is $10 \times 10$.

Algorithm: Per-region Background Model for Each Pixel
*For* the incoming local region $R_t$ centerd at curent pixel
1. Performe K-means clustering for pixels colors in region $R_t$.
2.Compute the means of each cluster's color $m_i$, where $i = 1, \dots, K$.
3. *For* $i = 1, \dots, K$.
Updating the per-region background model (GMM) $P_{region}$ with $m_i$.
*End for*
*End for*

**Fig.1.** The process of constructing per-region background model for each pixel.

Meanwhile, by performing above k-means clustering on pixels in the local region, we can get another 3-component GMM $P_{statistic}$, which describes statistic in the current local region. This 3-component GMM $P_{statistic}$ will be used in section 2.3 to adaptively adjust the learning rate $\alpha_r$ of the per-region background model to avoid false updating and maximize robustness.

### 2.3. Combining a Per-pixel and a Per-region Background Model in a Hierarchical Manner

Our algorithm combines a per-pixel with a per-region background model in the following hierarchical manner (see Fig.2). For each of those

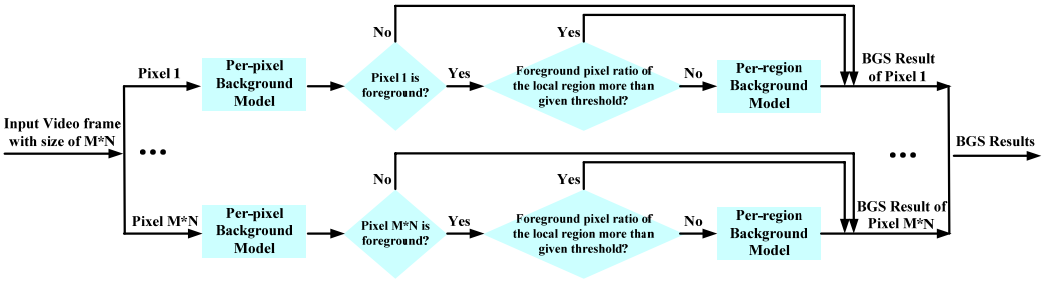


**Fig.2.** Overview of our method which combines a per-pixel with a per-region background model in a hierarchical manner.

pixels labeled as foreground by per-pixel processing, we determine whether it is a candidate false detection according to the foreground pixel ratio of the local region. If more than 50% of the pixels in a local region are labeled as foreground, we trust the current label. Otherwise, the current pixel is considered as a candidate false detection foreground pixel. Instead of using morphology or noise filtering to eliminate this kind of false detection, we use the per-region background model around this pixel to check if a match occurs. If the pixel matches the per-region background model, the pixel label is corrected as background. Otherwise, it is labeled as foreground. This is a natural combination because the two models have complementary strengths. The per-pixel background model is more precise than the per-region background model but is sensitive to noise and small movement of background. On the other hand, the per-region background model is less precise but more robust. Thus, our choice of combination of these two models accentuates the advantages of each.

In addition, in order to avoid false updating and maximize robustness, we adaptively adjust the learning rate $\alpha_r$ of a per-region background model as follows:

$$\alpha_r = \exp(-KL_{rs}/\gamma_{rs}) \quad (3)$$

where $\gamma_{rs}$ is a factor to control the influence of $KL_{rs}$, which is a similarity measure between the per-region background model $P_{region}$ and current local statistic model $P_{statistic}$. In this paper, we adopt an approximation of the Kullback-Liebler (KL) divergence between two GMM models [11]:

$$KL_{rs} = \sum_{k=0}^{K} \omega_k^r \min_i (KL(\eta_k^r || \eta_i^s) + \log \frac{\omega_k^r}{\omega_i^s}) \quad (4)$$

where $\eta_k^r$ and $\eta_i^s$ are the kth component of the per-region background GMM $P_{region}$ and the ith component of current local statistic GMM $P_{statistic}$ respectively. If the per-region background model and current local statistic model is similar, i.e., $KL_{rs}$ is small, the learning rate $\alpha_r$ is set to be large to adapt quickly. Otherwise, $\alpha_r$ is small (minimum value is 0.005).

## 3. Experiments and Discussion

The proposed method is implemented using C++, on a computer with Intel-Core 2 1.86 GHz processor. It achieves the processing speed of 10 fps at the resolution of $160 \times 120$ pixels. Identical parameters are used in the four sequences (see Fig. 3(a)). Even better performance could be achieved by adjusting the parameters for each video sequence.

We compare the performance of our method to the widely used method of GMM [1]. Both qualitative and quantitative comparisons are used to evaluate our approach. The quantitative comparison is done in terms of the number of false negatives (the number of foreground pixels that are missed) and false positives (the number of background pixels that are marked as foreground).

In Fig. 3(a), we show the results of the proposed method using four test sequences. The sequences used in the experiment include dynamic background, illumination changes and moderately camera jitter. The frames on the first column are from an outdoor sequence which contains a moving person in foreground, with dynamic background composed of subtle illumination variations in the sky along with ripples in the water and significant camera shake (being a hand-held camera). The next frames on the second column are from an outdoor sequence which contains a large amount of camera panning in the original video along with movement of trees with the wind. The proposed method robustly handles these situations and the moving object is detected correctly because it exploits spatial property of background image variations. The two sequences have been taken from [12]. The frames on the third column are from [7] where a foreground object is synthesized using similar color and with irregular shape. The proposed method also gives good results. The last frames on the fourth column are from [4]. The sequence contains average camera jitter of about 14.66 pixels. The proposed method does well under this condition.

In order to provide a quantitative perspective about the quality of foreground detection with our approach, we manually mark the foreground regions in five frames from each sequence to generate ground truth data, and make comparison with GMM. The numbers of error classifications are achieved by summing the errors from the frames corresponding to the ground truth frames. The corresponding quantitative comparison is reported in Fig. 3(b). For all sequences, the proposed method achieves best performance in terms of false positives, and false negatives are acceptable. Since the proposed method is obtained by combining a per-pixel background model with a per-region background model in a hierarchical manner, it is robust against dynamic background. It should be noticed that, for the proposed method, most of the false negatives occur on the contour areas of the foreground objects (see Fig.3 (a)). This is because the proposed method utilizes the spatial property of background image variations. According to the overall results, the proposed method outperforms the GMM for the used test sequences.

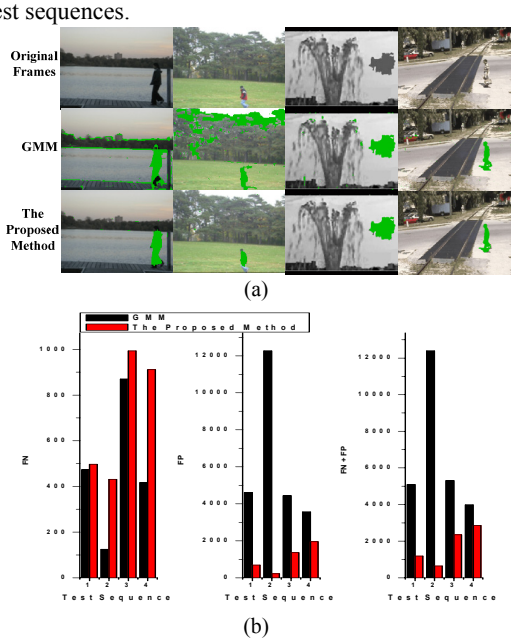


**Fig.3.** Comparision results of GMM and the proposed method. a) is the original test sequences and some detection results of the GMM and the proposed method. b) is the test results. FN and FP stand for false negatives and false positives, respectively.

## 4. Conclusion and Future Work

In this paper, we present a robust hierarchical background subtraction technique which takes advantages of meaningful correlation between pixels in the spatial vicinity to detect objects in difficult conditions. By combining a per-pixel with a per-region background model in a hierarchical manner, our method can precisely describe background change and tolerate variations in natural scenes, such as tree leaves swaying, water rippling, and camera jitter. Furthermore, we present a novel GMM based per-region background model to learn each pixel's local region background model. The proposed method has achieved lower false positives and better effectiveness, comparing to the widely used GMM.

Our future work will focus on how to cooperate our method with other features, such as the normalized RGB color of [3] and geometry to cope with the cast shadow, which is an extremely difficult problem in background subtraction.

## 5. Acknowledgements

This paper is partially supported by National Natural Science Foundation of China under contract No.60533030, No.60728203 and No.60775024.

## References

[1] C. Stauffer and W.E.L. Grimson. *Learning Patterns of Activity Using Real-Time Tracking*. TPAMI, vol. 22, no. 8, pp. 747-757, August 2000.

[2] O. Javed, K. Shafique, and M. Shah. *A Hierarchical Approach to Robust Background Subtraction using Color and Gradient Information*. IEEE Workshop on Motion and Video Computing, pp.22-27, 2002.

[3] A. Elgammal, D. Harwood, and L. Davis. *Non-parametric Model for Background Subtraction*. ECCV, vol.2, pp.751-767, June 2000.

[4] Y. Sheikh and M. Shah. *Bayesian Modeling of Dynamic Scenes for Object Detection*. TPAMI, vol. 27, no. 11, pp. 1778-1792, November 2005.

[5] K. Kim, T.H. Chalidabhongse, D. Harwood and L. Davis. *Real-time Foreground-Background Segmentation using Codebook Model*. Real-Time Imaging, vol.11, issue 3, pp.167-256, June 2005.

[6] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. *Wallflower: Principles and Practice of Background Maintenance*. ICCV, vol.1, pp.255-261, 1999.

[7] J. Zhong and S. Sclaroff. *Segmenting foreground objects from a dynamic, textured background via a robust kalman filter*. ICCV, vol.1, pp.44-50, 2003.

[8] Y.L. Tian, M. Lu, A. Hampapur. *Robust and efficient foreground analysis for real-time video surveillance*. CVPR, vol.1, pp.1182-1187, June 2005.

[9] K.A. Patwardhan, G. Sapiro and V. Morellas. *Robust Foreground Detection in Video Using Pixel Layers*. TPAMI, vol. 30, no. 4, pp. 746-751, April 2008.

[10] M.Piccardi. *Background subtraction techniques: a review*. SMC (4), pp. 3099-3104, 2004.

[11] J. Goldberger, S. Gordon, and H. Greenspan. *An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures*. ICCV, vol.1, pp.487 – 493, October 2003.

[12] http://www.tc.umn.edu/~patw0007/videolayer/